AN ABSTRACT OF THE THESIS OF

Mostafa Abdellatif Bedier for the degree of Doctor of Philosophy

in Statistics presented on September 18, 1989

Title: POST STRATIFIED ESTIMATION USING A KNOWN AUXILIARY VARIABLE

Abstract approved: Redacted for Privacy

G. David Faulkenberry

Post stratification is considered desirable in sample surveys for two reasons - it reduces the mean squared error when averaged over all possible samples, and it reduces the conditional bias when conditioned on stratum sample sizes. The problem studied in this thesis is post stratified estimation of a finite population mean when there is a known auxiliary variable for each population unit.

The primary direction of the thesis follows the lines of Holt and Smith (1979). A method is given for using the auxiliary variable in selection of the stratum boundaries and, using this approach to determine strata, to compare post stratified estimates with the self-weighting estimates from the analytical and empirical points of view. Estimates studied are: the post stratified mean, the post stratified combined ratio, and the post stratified separate ratio. The thesis contains simulation results that explore the distributions of the self-weighting estimates, and the post stratified estimates using conditional and unconditional inferences. The correct coverage

properties of the confidence intervals are compared and the design effect, i.e. the ratio of the variance of the self-weighting to the variance of post stratified estimates, is calculated from the samples and its distribution explored by the simulation study for several real and artificial populations. The confidence intervals of post stratified estimates using conditional variances had good coverage properties for each sample configuration used, and hence the correct coverage property over all possible samples provided that the Central Limit Theorem was applied.

The comparisons indicated that post stratification is an effective approach when the boundaries are obtained based on proper stratification using an auxiliary variable. Moreover it is more efficient than estimation based on simple random sampling in reducing the mean squared error.

Finally, there is strong evidence that the post stratified estimates are robust against poorly distributed samples, whereas empirical investigations suggested that the self-weighting estimates are very poor when the samples are unbalanced.

POST STRATIFIED ESTIMATION USING A KNOWN AUXILIARY VARIABLE

BY

MOSTAFA ABDELLATIF BEDIER

A THESIS

Submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Doctor of Philosophy

Completed **September 18, 1989**

Commencement **June 1990**

APPROVED:

Redacted for Privacy

Professor of Statistics in charge of major

Redacted for Privacy

Chairman of Department of Statistics

Redacted for Privacy

Dean Graduate School

Date thesis is presented September 18, 1989

ACKNOWLEDGMENTS

All Praise To Allah The Most Beneficent, The Most Merciful.

I would like to express my sincere appreciation and gratitude to my major professor Dr. David Faulkenberry for his valuable suggestions, guidance and excellent support throughout this work.

My sincere appreciation also to Dr. Lyle D. Calvin for being my on campus advisor in Dr. Faulkenberry's absence.

My gratitude to Dr. David R. Thomas and Dr. David S. Birkes for their assistance and suggestions for improving this work.

Many thanks to the faculty members, staff and students for their willingness to lend assistance throughout these studies.

Special thanks to brother Dr. Omar Dashwood for his assistance in edditing the thesis.

Finally, my sincere gratitude to my father and mother-in-law who died during the course of this Ph.D program, my mother for her praying for this work to be completed, and my wife, Souraya for her patience with my sons, Ahmed, Osama, Amir, and Ashraf.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES

POST STRATIFIED ESTIMATION USING A KNOWN AUXILIARY VARIABLE

1. INTRODUCTION

In finite population sampling designed based inference, it is well established that, when an auxiliary variable is available, proper use of stratified random sampling (STRS) reduces the variance of the estimate of the population mean compared to the variance of the estimate obtained from simple random sampling (SRS). When stratified random sampling is not used, but stratum information is available, a post stratified estimator may be used. In one aspect post stratification is potentially more efficient than stratification before selection, since after sampling the stratification factors can be chosen in different ways for different sets of variables in order to maximize the gains in precision.

Post stratification is considered desirable in sample surveys for two reasons ; it reduces the mean squared error when averaged over all possible samples, and it reduces the conditional bias when conditioned on stratum sample sizes.

It might be predicted that such a simple practical scheme would feature prominently in most texts on sampling, but this is not the case. As pointed out by Holt and Smith (1979) post stratification is rarely mentioned in text books, and its place in the literature may be described as 'modest' at best.

The situation of interest in this thesis is that where we take a simple random sample of fixed size n from a finite population of size N, and there is a known auxiliary variable, X_i, i = 1, . . . , N, for each population unit. The general problem considered is that of using the X_i in post stratification. This results in two problems. The first is how to choose the strata, and the second is the form of the post stratified estimator. Stratification based on the auxiliary variable requires specifying the sample allocation, choosing the number of strata, and determining boundaries for the strata. Obvious forms to consider for the estimator are the post stratified mean, the post stratified combined ratio estimator, and post stratified separate ratio estimator.

In this thesis a procedure for determining strata to use for post stratified estimation is given. Using this procedure, conditional and unconditional properties of the estimators are derived, and simulation is used to explore the distribution properties.

In addition to this introduction, this thesis contains eight other chapters, whose contents are as follows.

Chapter 2 discusses the problems, notation and previous solutions of post stratification. A few applications are briefly introduced, employing either unconditional inferences such as those of Williams (1962) and Fuller (1966), or conditional inferences based on Holt and Smith (1979). The chapter concludes with a description of the population used in the simulation studies.

In chapter 3 the notation of " optimum stratification" is introduced

and the boundaries are derived using the equations of Dalenius (1950). In addition to Dalenius's work, the appropriate choice of boundaries based on the auxiliary variable is introduced, and comparisons are made between the post stratified estimator obtained by this approach and that obtained based on categorical boundaries under proportional allocation.

Chapter 4 discusses the distribution of the self-weighting mean averaging over all possible allocations, or conditioning on the sample allocation, and its conditional bias is derived and explored by a computer simulation for several real and artificial populations.

In chapter 5 the distribution of the post stratified mean using conditional inference for different sample sizes is introduced, averaging over all possible sample configurations. In this chapter the distribution of post stratified mean conditioning on the sample configuration is also studied under three different configurations. Chapter 5 concludes with a comparison of post stratified and self-weighting means, where the comparisons are made analytically and empirically for the populations under study.

Chapter 6 focuses on the distribution of the ratio estimate, and its conditional bias and the leading term of the bias will be introduced. In this chapter a simulation study is made to explore the distribution of the self-weighting ratio estimate averaging over all possible sample allocations. The distribution of self-weighting ratio estimate conditioned on a particular sample allocation also is introduced, and the chapter concludes with a comparison of the self-weighting ratio estimate and the post stratified mean for the populations under study.

In chapter 7 the distribution of the post stratified combined ratio

estimator is introduced and compared to the self-weighting ratio estimator. The distribution of the post stratified combined ratio estimator conditioned on the sample allocation is discussed and sample variance obtained. A computer simulation will be used in this chapter to explore the distribution of the combined ratio estimator averaging over all possible sample allocations. The distribution of the combined ratio estimator conditioned on the sample allocation is also introduced. The chapter ends with analytical and empirical comparisons between post stratified combined and self-weighting ratio estimates.

Chapter 8 discusses the distribution of the post stratified separate ratio estimator and compares it with the self-weighting ratio estimator. The distribution of the post stratified separate ratio estimator conditioned on the sample allocation is highlighted and sample variance obtained. A computer simulation is used to explore the distribution of the separate ratio estimator averaging over all possible sample allocations. The distribution of the separate ratio estimator conditioned on the sample allocation is also introduced. The chapter concludes with analytical and empirical comparisons between post stratified separate and self-weighting ratio estimates.

Finally, chapter 9 consists of an overview and final conclusions of the entire work presented in this thesis.

2. REVIEW OF LITERATURE

To improve the quality of estimates in sample surveys some kind of weighting is often carried out. Post stratification is a popular weighting method and is considered desirable for two reasons :

1. It reduces the mean squared error (MSE) when averaged over all possible samples.
2. It reduces conditional bias when conditioned on stratum sample size.

Post stratification is potentially more efficient than stratification before sample selection, since after sampling the stratification factors can be chosen in different ways for different sets of variables in order to maximize the gains in precision. One would have thought that such a simple practical scheme would feature prominent in most texts on sampling, but this is not the case. As pointed by Holt and Smith (1979) it is seldom mentioned in text books, and its place in the literature is modest.

In this chapter the post stratification concept is formulated, and previous work dealing with methods 1 and 2 above is discussed.

2.1. The Problems and Notation

Suppose that the population under study comprises N units 1, . . ., N. Associated with unit i is an unknown variable Y_i and a known auxiliary variable X_i . Assume the population can be partitioned into L strata of sizes $N_1, \ldots, N_L, \sum_{h=1}^{L} N_h = N, h=1, \ldots, L$.
The stratum weights, means and variances are :

$W_h = N_h / N$,

$$\bar{Y} = \sum_{h=1}^{L} \sum_{j=1}^{N_h} Y_{hi} / N, \quad S^2 = \sum_{h=1}^{L} \sum_{j=1}^{N_h} (Y_{hj} - \bar{Y})^2 / (N - 1),$$

$$\bar{Y}_h = \sum_{j} Y_{hj} / N_h, \quad S_h^2 = \sum_{j} (Y_{hj} - \bar{Y}_h)^2 / (N_h - 1),$$

A sample of fixed size n is taken, which after selection falls into the strata according to the vector $n = (n_1, \ldots, n_L)$, $\sum_{h=1}^{L} n_h = n$.
The components of n are not known until after the sample is drawn.
The sample yields values y_{hj}, $h = 1, \ldots, L$, $j \in s$, where s denotes the set of units in the sample.
The sample proportion, mean and variance in the h_{th} stratum are:

$w_h = n_h / n$,

$$\bar{y}_h = \sum_{j} y_{hi} / n_h, \quad s_h^2 = \sum_{j}^{n_h} (y_{hj} - \bar{y}_h)^2 / (n_h - 1)$$

The finite population correction factors are $f = n / N$ and $f_h = n_h / N_h$.
The post stratified estimator of $\bar{Y}$, assumes N_h known is given by

$$\bar{y}_{ps} = \sum_{h} W_h \bar{y}_h \qquad (2.1)$$

It is clear that each stratum mean is weighted by the stratum proportion.
Thus if a sample is badly balanced for some characteristic the post stratified estimator automatically corrects for this.

The variance of $\bar{y}_{ps}$ depends on the sampling distribution to which it

is related . There are two methods of calculating the variance of $\bar{y}_{ps}$. First, the variance can be determined by the distribution conditional on the vector n of stratum sample sizes actually attained in the sample under study. Second, the variance can be determined by the unconditional distribution determined by all possible samples of fixed size n.

The conditional variance of $\bar{y}_{ps}$ is simply the usual variance for stratified samples given by

$$V\{ (\bar{y}_{ps} \mid n) \} = \sum_{h} W_h^2 (1 - f_h) S_h^2 / n_h \qquad (2.2)$$

The unconditional variance is obtained by averaging (2.2) over all possible distributions of n . This gives

$$V(\bar{y}_{ps}) = \sum_{h} W_h^2 S_h^2 \{ E(n_h^{-1}) - N_h^{-1} \}$$

$$\doteq \{ (1 - f) / n \} \sum_{h} W_h S_h^2 + \sum_{h} (1 - W_h) S_h^2 / n^2 \qquad (2.3)$$

where $E(n_h^{-1})$ is obtained by the method of Stephen (1945), and assumes $n_h > 0$ for all h and is given by

$$E(n_h^{-1}) = (1 / nW_h) + (1 - W_h) / n^2 W_h^2 \qquad (2.4)$$

If $n_h = 0$ for some h then $\bar{y}_{ps}$ is not defined and neither variance can be employed directly. One practical solution is to pool or collapse similar strata.

Some authors advocate employing the unconditional variance (2.3) including Hansen et al. (1953), Des Raj (1972), Cochran (1963),

and Kish (1965) with S_h^2 estimated by s_h^2. Others advocate the conditional variance (2.2), such as Yates (1960) , with S_h^2 estimated by s_h^2 . Durbin (1969) argued that the achieved sample size should be treated as an ancillary statistic. Cox and Hinkely (1974) adopt a similar approach Holt and Smith (1979) argued that the inferences should be made conditional on the achieved sample configuration.

Kalton in his discussion of Smith (1984) argued in favor of conditional inferences with S_h^2 estimated by s_h^2 . If there is an unbiased estimate of the conditional variance, it is also unbiased for its unconditional variance.

Rao (1985) argued that the inference should be conditional on the observed sample sizes if the sample sizes are random and their population distributions is known.

Fuller (1966) proposed an alternative procedure where the unconditional framework is usually superior to that of collapsing empty strata.

Williams (1962) introduced a method for obtaining the unconditional variance of a post stratified estimator in any type of sampling .

If the stratum boundaries are not suggested a priori, the variance of $\bar{y}_{ps}$ for a given number of strata and a sample configuration is a function of the location of the stratum demarcation points, whose selection specifies the values of W_h and S_h^2.

2.2 Previous Solutions

2.2.1 An approximate formula for a post stratification variance

Williams (1962) presented a method for obtaining the unconditional variance of a post stratified estimator in any type of sampling. His approach was to use the variance of the ratio estimator to obtain an approximation for the variance of a post stratified estimator.

Consider a sample of size n that has been drawn according to any specified sampling scheme from a population of size N. Let $\hat{Y} = \hat{Y}(y)$ denote an estimator of the population total, $V(\hat{Y}) = \sigma_{\hat{Y}}^2(y)$ the sampling variance of $\hat{Y}$, and $V(\hat{Y}) = \hat{\sigma}^2_{\hat{Y}}(y)$ an estimator of $V(\hat{Y})$.
Williams (1962) defined the following

$y' = y$ if the unit is in the h_{th} stratum,

$= 0$ if the unit is not in the h_{th} stratum

and

$c = 1$ if the unit is in the h_{th} stratum,

$= 0$ if the unit is not in the h_{th} stratum,

then $\hat{N}_h = \hat{Y}(c)$, $\hat{Y}_h = \hat{Y}(y')$ are respectively, an estimator of the h_{th} stratum size, and the stratum total.
Thus $\hat{y}_h = \hat{Y}_h / \hat{N}_h$ is a ratio estimator of the stratum mean with sampling approximately given by

$V(\hat{y}_h) = \sigma^2_{\hat{Y}}(w') / N_h^2$

where $w' = y - \bar{Y}_h$ if a unit is in the t_{th} stratum,

$= 0$ if it is not.

Then the post stratified weighting estimator $\hat{y}_p$ is given by

$$\hat{y}_p = \sum_h W_h \hat{y}_h \qquad (2.5)$$

with variance

$$V(\hat{y}_p) = \sigma^2_{\hat{Y}}(w) / N^2 \qquad (2.6)$$

where $w = y - \bar{Y}_h$ if a unit is in the h_{th} stratum,

$= 0$ if a unit is not,

$h = 1, 2, \ldots, L.$

Thus in order to obtain the variance of $\hat{y}_p$ in any type of sampling insert the variant, w into the formula $\sigma^2_{\hat{Y}}(y)$ in equation (2.6) of the specific design and divide by N^2.

This is the main result obtained by Williams (1962) in the framework of unconditional inferences for the post stratification of many types of samples. This result should be used with caution when small sample sizes are employed, since the post stratified estimator is assumed to be a ratio estimator.

2.2.2 An unbiased post stratified estimator admitting empty strata

Certain practical difficulties are associated with the fact that the size of the sample falling in any particular stratum is a random variable . Thus, one should either use strata large enough to reduce the probability of zero sample size or use an estimation procedure for empty strata that has biases of unknown magnitude. Fuller (1966) developed small sample estimators for two post strata and compared with pooling or collapsing procedures commonly employed in practice. The estimators are not necessarily conditionally unbiased for a particular sample configuration, but their conditional MSE might be smaller than that of the common post stratified estimator.
The estimation for two post strata shall be briefly stated. Given a SRS of

size n from a population known to contain W_1 and $W_2 = 1 - W_1$ population proportions, the post stratified estimator of the mean is given by

$$\bar{y}_p = A_i \bar{y}_1 + 1 - (A_i) \bar{y}_2 \qquad (2.7)$$

where $A_i \in [0, 1]$ is the weight applied to the mean of stratum 1 for samples with $i = 0, 1, 2, \ldots, n$ sample elements in stratum one. Fuller (1966) also defined

$$A_0 = 0$$

$$A_n = 1$$

This indicates that the sample mean is used as the estimator if one of the strata contains no sample elements. The conditional MSE for the estimator given i sample elements in stratum one is given by

$$V(\bar{y}_p) = A_i^2 f_{1i} \left(\frac{1}{i}\right) S_1^2 + (1 - A_i) f_{2i} \left(\frac{1}{n - i}\right) S_2^2 + (A_i - W_1)^2 (\bar{Y}_1 - \bar{Y}_2)^2 \qquad (2.8)$$

where $A_i \left(\frac{1}{i}\right)$, $(1 - A_i)^2 \left(\frac{1}{n - i}\right)$ are defined to be zero if $i = 0$ or $i = n$ respectively,

$$f_{1i} = \frac{N_2 - n + i}{N_2},$$

and A_i is obtained by minimizing (2.8) with respect to it. This gives

$$A_i = \frac{i f_{2i} S_2^2 + i(1 - i) W_1 (\bar{Y}_1 - \bar{Y}_2)^2}{(n - i) f_{1i} S_1^2 + i f_{2i} S_2^2 + i(n - i)(\bar{Y}_1 - \bar{Y}_2)^2} \qquad (2.9)$$

This estimator (2.7) is unbiased under the condition

$$E(\bar{y}_h \mid n_h) = \bar{Y}_h \quad , n_h > 0.$$

Fuller (1966) also generalized the two strata procedures using unequal

probability sampling to any number of strata, where the population is first divided into two groups and the groups are repeatedly divided into groups of two. At the first subdivision, an unbiased estimator can be constructed for each pair of strata. An example for five strata was given in the paper of Fuller (1966), but the formula for the weights is very complicated thus making it difficult for practical use.

2.2.3 Post stratification as a robust technique

Holt and Smith (1979) argued that inferences should be made conditional on the achieved sample configuration n . The unconditional expected values and variance could be used at the design stage before the sample is drawn.

These authors compared the post stratified estimate $\bar{y}_{ps}$ defined in (2.1) with the self-weighting estimator $\bar{y}_{sw}$ with a simple random sample (SRS) . This estimator can be defined as

$$\bar{y}_{sw} = \sum_{h} \sum_{j \epsilon s} y_{hj} / n = \sum_{h} w_h \bar{y}_h \qquad (2.10)$$

and conditional on n the conditional expectation is given by

$$E\{(\bar{y}_{sw} \mid n)\} = \bar{Y} - \sum_{h} \bar{Y}_h (W_h - w_h) \qquad (2.11)$$

The last expression is the conditional bias. Thus the conditional MSE of $\bar{y}_{sw}$ is given by

$$MSE\{(\bar{y}_{sw} \mid n)\} = \sum_{h} w_h^2 (1 - f_h) \frac{S_h^2}{n_h} + \left\{ \sum_{h} \bar{Y}_h (W_h - w_h) \right\}^2 \qquad (2.12)$$

Comparing the MSE in (2.12) with the variance of $\bar{y}_{ps}$ in (2.2) gives

$$MSE\{(\bar{y}_{sw} \mid n)\} - V\{(\bar{y}_{ps} \mid n)\} = \left\{\sum_h \bar{Y}_h (W_h - w_h)\right\}^2 + \sum_h \left\{w_h^2 - W_h^2\right\}(1 - f_h)\frac{S_h^2}{n_h} \qquad (2.13)$$

The difference in (2.13) is either positive or negative depending on the sample configuration n relative to the post strata mean and variances. Hence, Holt and Smith (1979) could not say that one estimator is uniformly better than the other, but their empirical investigation indicated that post stratification offers protection against unfavorable sample configurations and should be viewed as a robust technique.

2.3. Populations Used for Empirical Study

In this section four varieties of real and artificial populations are introduced to compare self-weighting and post stratified estimators. Two populations are real in the sense that genuine data have been used to obtain the values for the stratum means, populations and variance. These two populations are obtained from an agricultural survey. The other populations are generated from linear models. Further details of which are given in Table 2.1.

Table 2.1. *Populations used in empirical study*

Case	Strata	Description
Population source: Model Generated		
1.	3	$y = c + x + c z$, where $c = .1$ x has chi square distribution with one degree of freedom and z has standard normal distribution.
2.	3	y is generated from the above model where x has chi square distribution with three degree of freedom.

Case	Strata	Study variable	Stratification variable
Population source: Agricultural Survey			
3.	4	Corn production	Farm size
4.	4	Wheat production	Farm size

3. THE USE OF AN AUXILIARY VARIABLE IN POST STRATIFICATION

When the relationship between the study and auxiliary variables is linear one approach to determining the stratum boundaries is to employ Dalenius equations on the auxiliary variable . In this chapter, the problem of optimum post stratification is presented under the condition that stratification is carried out on a known auxiliary variable. Our interest is primarily concerned with whether this approach provides stratification offering reasonable improvements over stratification based on arbitrary or categorical boundaries, and whether it is more efficient than SRS.

In the next section, optimum stratification is introduced and the best boundaries are derived using Dalenius and Gurney equations.
In the third section, the choice of strata boundaries based on the auxiliary variable using Dalenius's original equations is introduced.
Comparisons will then be made between the post stratified estimator obtained by this approach and that obtained based on categorical boundaries. Since stratification is employed after taking the SRS, the choice of boundaries is made under proportional allocation which is the expected allocation.

3.1. Optimum Stratification

The term " optimum stratification " refers to choosing those stratum boundaries that minimize the variance of the stratified random sampling estimate of the population mean when the number of strata and the method of allocation are fixed. The problem was first examined theoretically by Dalenius (1950), and his method shall be considered briefly. The theories

of sample surveys are based on sampling from a finite population.

Since this approach involves the use of detailed algebra, we will begin by supposing that the study population' s distribution can be represented by the probability distribution function, $f(y)$, $a \leq y \leq b$.

If $a < y_1 < y_2 < \ldots < y_{L-1} < b$ are the points of demarcation between the L strata, then the mean and the variance of the study variable within the hth stratum can be defined by

$$W_h = \int_{y_{h-1}}^{y_h} f(t)\, dt, \quad W_h\mu_h = \int_{y_{h-1}}^{y_h} t\, f(t)\, dt,$$
$$W_h\sigma_h^2 = \int_{y_{h-1}}^{y_h} t^2 f(t)\, dt - W_h\mu_h^2 \qquad (3.1)$$

for h = 1, 2, . . . , L. With $f(y)$ known and the number of strata considered fixed, the variance of $\hat{\mu}$ for proportion allocation , $V((\hat{\mu} \mid \text{prop}))$, is a function of the stratum boundaries only. If the finite population correction factor is ignored this gives

$$V((\hat{\mu} \mid \text{prop})) = \sum_h W_h\, \sigma_h^2 / n \qquad (3.2)$$

Thus , to determine the best boundaries y_h that provide the greatest reduction in (3.2) differentiate $\sum W_h\sigma_h^2$ with respect to y_h and equate the expression thus obtained to zero

$$\frac{\partial V((\hat{\mu} \mid \text{prop}))}{\partial y_h} = 0 \qquad h = 1, 2, \ldots, L\text{-}1.$$

After some manipulation (see Des raj (1972) page. 70) this gives

$$y_h = \frac{1}{2}(\mu_h + \mu_{h+1}) \qquad (3.3)$$

This shows that the best y_h value is the average of the two strata means which it separates. The problem is that y_h will have to be found iteratively, since the value of μ_h depends on y_h and this is difficult to solve. Other values for y_h are obtained for optimum and equal allocations but we will only study the proportion allocation since post stratification closely resembles this allocation .

Dalenius and Gurney (1951) considered the problem of optimum stratification when stratification was carried out on a specific auxiliary variable. They assumed that the two variables were related by the relationship $y = g(x) + e$, where X is the auxiliary variable with probability distribution function $f(x)$, $a' \leq x \leq b'$, and where e is the random deviation of Y from the regression line $y = g(x)$, X and e are uncorrelated, $E(e) = 0$. The moments of Y within the strata can be expressed as a function of the stratum moments of $g(x)$ and X, thus making the variance , $V((\hat{\mu} \mid \text{prop}))$ a function of the stratum boundaries with respect to the auxiliary variable, $a' < x_1 < x_2 < \ldots < x_{L-1} < b'$. Using the model obtained by Singh and Sukhatme (1969), the following may be defined

$$W_h = \int_{x_{h-1}}^{x_h} f(t)\, dt,$$

$$W_h \, E[\, g(X) \mid h \,] = \int_{x_{h-1}}^{x_h} g(t)\, f(t)\, dt$$

$$\sigma_h^2 = E[\, g(X)^2 \mid h \,] - E[\, g(X) \mid h \,]^2$$

and $$W_h \, \sigma_{he}^2 = \int_{x_{h-1}}^{x_h} \phi(t)\, f(t)\, dt = E[\, \phi\,(X) \mid h \,], \quad \phi\,(x) = \text{var}[\, e \mid x \,]$$

with $h = 1, 2, \ldots, L$. Substituting these functions into equation (3.2) the boundaries of the optimum stratification are the solutions of

$$\frac{\partial V[(\, \hat{\mu} \mid \text{prop}\,)]}{\partial x_h} = 0 \qquad h = 1, 2, \ldots, L\text{-}1.$$

The resulting equation is given by

$$g\,(x_h) = \frac{E[\, g\,(X) \mid h \,] + E[\, g(X) \mid h{+}1]}{2} \tag{3.4}$$

As was true with equation (3.3) the Dalenius-Gurney equation (3.4) must be solved iteratively, and this calculation can rather difficult. For this reason, approximate rules have been found in which the variance can be reduced directly. The most practical approach, in terms of balancing ease of solution and at the same time including the most information about the functional relationship between the study and auxiliary variables, seems to be the cum H(x) rules, i.e. those equations of the form

$$\int_{x_{h-1}}^{x_h} H\,(t)\, dt = \text{constant} = \frac{\int_{a'}^{b'} H\,(t)\, dt}{L} \qquad h = 1, 2, \ldots, L\text{-}1. \tag{3.5}$$

where the boundaries that solve these equations equalize the cumulative affect of the function H (x). Thomsen (1976) obtained an approximation to Dalenius' s equation having this H (x) rule form, that is, cum $\sqrt[3]{f(x)}$ rule for proportional allocation. Singh (1975) generated an analogous cum H(x) under proportional allocation given by

$$H(x) = \sqrt[3]{\{f(x)\,[\,g'(x) + \theta\,\phi(x)]\}}\,, \quad \theta = \frac{12\,L^2}{(b' - a')^2} \tag{3.6}$$

Chester (1980) showed that this cum H(x) in (3.6) was the best rule for both linear and nonlinear relationships between Y and X, that can be reduced to $\sqrt[3]{f(x)}$ when the relationship is just simple linear.
It was also established (Chester, 1980) that the most efficient number of strata usually lies between three and eight strata.

3.2 The Choice of Stratum Boundaries Based on the Auxiliary Variable

Suppose that Y and X are linearly correlated, X has the distribution function $f(x)$, $b_0 \le x \le b_L$ and $f(y\,,\,x)$ is the joint distribution, and $E(\,Y \mid h\,) = E(\,Y \mid b_{h-1} \le x \le b_h\,)$.
Thus the variance of $\bar{y}_{ps}$ defined in (3.2) is a function of stratum boundaries . For comparison (3.2) can be written in the form

$$V\{\,(\,\bar{y}_{ps(1)} \mid \text{prop}\,)\,\} = \frac{1}{n}\sum_{h} W_h\, S^2_{h(1)} \tag{3.7}$$

where a finite population correction factor is ignored, and the subscript (1) refers to the stratification based on arbitrary boundaries for proportional

allocation.

Define

$$W_h = \int_{b_{h-1}}^{b_h} f(x)\,dx \qquad \text{and} \qquad \frac{\partial W_h}{\partial b_h} = f(b_h)\,,$$

and the density of Y within stratum h may be written as

$$g_h(y) = \frac{1}{W_h} \int_{b_{h-1}}^{b_h} f(y \mid x)\, f(x)\,dx.$$

The mean and variance of Y within stratum h are given by

$$\mu_{yh} = \int_Y y\, g_h(y)\,dy\,,$$

$$S_{yh}^2 = \int_Y (y - \mu_{yh})^2\, g_h(y)\,dy = \int_Y y^2\, g_h(y)\,dy - \left\{\int_Y y\, g_h(y)\,dy\right\}^2$$

$$W_h S_{yh}^2 = \int_Y y^2 W_h\, g_h(y)\,dy - \frac{\left\{\int_Y y\, W_h\, g_h(y)\,dy\right\}^2}{\int_{b_{h-1}}^{b_h} f(x)\,dx} \tag{3.8}$$

By substituting the above equations into (3.7), the best boundaries b_h that provide the greatest reduction in this variance are the solution of the equations:

$$\frac{\partial V\{(\bar{y}_{ps(1)} \mid \text{prop})\}}{\partial b_h} = 0\,, \quad h = 1,\ 2,\ \ldots,\ L\text{-}1$$

This is equivalent to minimizing $\sum_h W_h S_h^2$ with respect to b_h, and is

sufficient to differentiate $W_h S_h^2 + W_{h+1} S_{h+1}^2$ with respect to b_h.

From equation (3.8) we have

$$\frac{\partial W_h S_h^2}{\partial b_h} = \frac{\partial}{b_h}\left\{ \int_{b_{h-1}}^{b_h} E(Y^2 \mid X)\, f(x)\, dx - \frac{\left(\int_{b_{h-1}}^{b_h} E(Y \mid X)\, f(x)\, dx \right)^2}{\int_{b_{h-1}}^{b_h} f(x)\, dx} \right\}$$

$$= E(Y^2 \mid X = b_h) f(b_h) - 2\, E(Y \mid h)\, E(Y \mid X = b_h) f(b_h) + E^2(Y \mid h) f(b_h)$$

Similarly ,

$$\frac{\partial W_{h+1} S_{h+1}^2}{\partial b_h} = - E(Y^2 \mid X = b_h) f(b_h)$$

$$+ 2\, E(Y \mid h+1)\, E(Y \mid X = b_h) f(b_h) - E^2(Y \mid h+1) f(b_h)$$

Thus , $\frac{\partial}{\partial b_h} \{ W_h S_h^2 + W_{h+1} S_{h+1}^2 \} = 0$, gives

$$2 \left[E(Y \mid X = b_h) \right] \left[E(Y \mid h+1) - E(Y \mid h) \right] = E^2(Y \mid h+1) - E^2(Y \mid h)$$

This implies

$$E(Y \mid X = b_h) = \frac{E(Y \mid h) + E(Y \mid h+1)}{2} \qquad (3.9)$$

Since Y is linearly correlated to X , then (3.9) is true if

$$b_h = \frac{E(X \mid h) + E(X \mid h+1)}{2} \qquad (3.10)$$

where h = 1, 2, . . . , L-1. This result in (3.10) coincides with the equation (3.4) obtained by Dalenius and Gurney (1951) using the model of Singh and Sukhatmes (1969).

Certainly, if one uses these boundaries for stratification, the post stratified estimator $\bar{y}_{ps(2)}$ would have less variance than the variance of $\bar{y}_{ps(1)}$. However, as was the case with equation (3.4), equation (3.10) must be solved iteratively since the means depend on the boundaries.

An approximation method, due to Singh (1975) and Thomsen (1976) to minimize (3.7) is as follows:

let $$Z(x) = \int_{b_0}^{b_L} \sqrt[3]{f(x)}\, dx$$

If the strata are numerous and narrow, $f(x)$ should be approximately constant within a given stratum. Hence

$$W_h = \int_{b_{h-1}}^{b_h} f(x)\, dx \doteq f_h(b_h - b_{h-1}),$$

$$Z_h - Z_{h-1} = \int_{b_{h-1}}^{b_h} \sqrt[3]{f(x)}\, dx \doteq \sqrt[3]{f_h}\,(b_h - b_{h-1})$$

where f_h is a "constant" value of $f(x)$ in stratum h. Substituting these equations in (3.7) we have

$$12 \sum_h W_h S_h^2 = \sum_h f_h (b_h - b_{h-1})^3 \doteq \sum_h (Z_h - Z_{h-1})^3$$

Since $(Z_h - Z_{h-1})$ is fixed, the sum to the right of the equation is minimized by making $(Z_h - Z_{h-1})$ constant.

Given $f(x)$, to construct strata, the rule is to form the cumulative value of $\sqrt[3]{f(x)}$ and choose the x_h so that they create equal intervals on the cum $\sqrt[3]{f(x)}$ scale.

From the above argument it can be concluded that the post stratified estimate based on optimum boundaries is better than that obtained based on arbitrary boundaries, so that in all following sections the above cum $\sqrt[3]{f(x)}$ approximation will be used to estimate the best boundaries that form the strata.

4. DISTRIBUTION OF THE SELF-WEIGHTING MEAN

The self-weighting estimator $\bar{y}_{sw}$ defined in (2.10) is known to be unbiased under SRS with variance

$$V(\bar{y}_{sw}) = (1 - f) S^2 / n \qquad (4.1)$$

However $\bar{y}_{sw}$ is a biased estimator conditional on the sample allocation. The conditional expectation and MSE of $\bar{y}_{sw}$ are given in (2.11) and (2.12) respectively.

Hence, the conditional bias of $\bar{y}_{sw}$ is given by

$$E\{(\bar{y}_{sw} \mid n)\} - \bar{Y} = \sum_h \bar{Y}_h (w_h - W_h) \qquad (4.2)$$

The bias is zero when the sample is proportionally distributed over strata, that is, when $w_h = W_h$. The bias is also zero when $\bar{Y}_h$ is constant for all h. In other cases the estimator is conditionally biased. However, when referred to the unconditional sampling distribution, $\bar{y}_{sw}$ is always deemed to be unbiased even for samples with highly disproportionate allocations across post strata.

4.1. The Distribution of $\bar{y}_{sw}$

For each population under study given in Table 2.1 with a known auxiliary variable, stratum boundaries will be determined by the cum $\sqrt[3]{f(x)}$ approximation rule, and stratum means and variances will be estimated from the sample. For any given sample allocation n, calculate the sample variance $s^2_{\bar{y}_{sw}}$ of $\bar{y}_{sw}$ that has the usual form

$$s^2_{\bar{y}_{sw}} = (1 - f) s^2 / n \qquad (4.3)$$

and s^2 is the SRS sample variance.

A computer simulation was used to explore the distributions of

$$t_{sw} = (\bar{y}_{sw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{sw}}} \qquad (4.4)$$

The achieved sample allocation may be obtained by generating one uniform deviate on the range (0, 1) for each member of the sample. These random numbers were then used to determine from which stratum each individual was drawn and so gave n ; the value of t_{sw} then can be calculated.
The entire process was repeated 1000 times for each population and for each total sample size, and the distributions of t_{sw} were explored.
To avoid any problems a sample allocation of one or zero in any stratum was rejected, but the total sample size and stratum proportions in each case were such as to make this extremely unlikely. Thus, the overall impression of the results obtained should essentially remain unaffected.

Table 4.1 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th

percentiles of t_{sw} for the four cases under study (see Table 2.1).

Table 4.1 shows that t_{sw} has a distribution similar to the Z distribution in the two artificial cases 1 and 2, but it is not similar to Z in the other two cases (3 and 4) when sample sizes are relatively small.

4.2 The Distribution of $\bar{y}_{sw}$ Conditioned On The Sample Configuration

We have argued that the self-weighting mean is conditionally biased, and the conditional bias is defined in (4.2). In this section another simulation study is made to explore the distribution of $\bar{y}_{sw}$ conditioning on a particular sample allocation. For each population in Table 2.1 the stratum proportions are used as a set of multinomial probabilities, assuming large strata, and the achieved sample allocation is obtained by repeating the sampling process until the required allocation n is obtained. The t_{sw} can then be calculated. The entire process was repeated 1000 times for each population and for three allocations, namely proportional allocation (prop), disproportional but sample units are allocated relative to stratum size (propn), and disproportional allocation (propd). The latter allocations were chosen in order to study the performance of the estimates when the sample allocation is altered from the proportion allocation. For example, in case 1 the population is post stratified into three strata with sample proportion, $f_h = n_h/N_h$ was chosen such that: f_1= 0.05 0.05 0.05, f_2 = 0.03 0.20 0.26, and f_3 = 0.01 0.20 0.65 under prop, propn, and propd respectively. In case 2 f_h was chosen such that: f_1 = 0.05 0.05 0.05, f_2 = 0.04 0.05 0.14, and f_3 = 0.02 0.08 0.21 respectively. For cases 3 and 4 where the populations were post stratified into four strata, the sample proportion f_h was chosen such

that: f_1 = 0.04 0.04 0.04 0.04, f_2 = 0.02 0.05 0.05 0.23, f_3 = 0.02 0.07 0.02 0.45 under prop, propn, and propd respectively.

Table 4.2 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{sw} for the four cases under study and under three allocations.

It is evident from Table 4.2 that t_{sw} has a distribution similar to the Z distribution under the prop allocation, but under the propn and propd allocations, percentiles indicate that t_{sw} has only positive percentiles, suggesting that $\bar{y}_{SW}$ is biased upward.

It is a good idea to plot percentiles of the distribution of t_{sw} against the percentiles of Z distribution that are presented for comparative purposes in Table 4.2. Figures 4.1- 4.4 clearly show the conditional bias of $\bar{y}_{SW}$, in particular when the sample allocation is altered from the proportional allocation.

Table 4.1. *Percentile of the distribution of*

$t_{sw} = (\bar{y}_{sw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{sw}}}$, and $t_{ps} = (\bar{y}_{ps} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{psln}}}$

Percentiles of	t_{sw}			t_{ps}			Z
	Sample Size			Sample Size			
Case 1.	100	200	300	100	200	300	
1%	-2.40	-3.13	-2.66	-2.79	-2.87	-2.73	-2.33
5%	-1.71	-1.94	-1.87	-1.87	-2.00	-1.94	-1.65
10%	-1.28	-1.46	-1.45	-1.58	-1.54	-1.53	-1.28
20%	-0.78	-0.94	-0.92	-1.00	-0.93	-0.96	-0.84
30%	-0.45	-0.57	-0.58	-0.63	-0.61	-0.64	-0.52
40%	-0.17	-0.25	-0.28	-0.31	-0.28	-0.34	-0.25
50%	0.09	-0.01	-0.01	-0.03	-0.05	-0.04	0.0
60%	0.29	0.25	0.24	0.22	0.19	0.26	0.25
70%	0.54	0.50	0.49	0.48	0.48	0.54	0.52
80%	0.83	0.82	0.81	0.85	0.80	0.84	0.84
90%	1.24	1.18	1.17	1.30	1.24	1.24	1.28
95%	1.50	1.45	1.53	1.67	1.49	1.64	1.65
99%	2.07	2.08	2.20	2.25	2.31	2.15	2.33
Case 2.							
1%	-2.87	-2.73	-2.36	-2.42	-2.22	-2.17	
5%	-1.81	-1.75	-1.72	-1.62	-1.59	-1.69	
10%	-1.32	-1.30	-1.31	-1.19	-1.24	-1.27	
20%	-0.81	-0.85	-0.83	-0.78	-0.76	-0.81	
30%	-0.49	-0.45	-0.46	-0.43	-0.44	-0.53	
40%	-0.25	-0.22	-0.20	-0.19	-0.19	-0.23	
50%	0.03	-0.04	0.02	0.06	0.07	-0.03	
60%	0.27	0.29	0.25	0.28	0.29	0.22	
70%	0.51	0.52	0.50	0.54	0.51	0.50	
80%	0.79	0.80	0.83	0.85	0.81	0.75	
90%	1.21	1.17	1.22	1.25	1.21	1.19	
95%	1.54	1.45	1.49	1.66	1.54	1.60	
99%	2.19	2.03	1.99	2.27	2.32	2.16	

Table 4.1. Continued

Percentiles of	t_{sw}			t_{ps}			Z
	Sample Size			Sample Size			
Case 3.	100	200	300	100	200	300	
1%	-2.14	-2.58	-2.91	-2.87	-3.10	-2.99	
5%	-1.35	-1.71	-1.97	-1.99	-2.03	-1.93	
10%	-0.94	-1.38	-1.53	-1.50	-1.49	-1.41	
20%	-0.50	-0.89	-0.99	-1.00	-0.94	-0.89	
30%	-0.17	-0.54	-0.62	-0.59	-0.59	-0.56	
40%	0.09	-0.25	-0.33	-0.29	-0.32	-0.28	
50%	0.35	0.02	-0.08	-0.03	-0.08	-0.02	
60%	0.53	0.27	0.17	0.19	0.15	0.24	
70%	0.75	0.55	0.42	0.48	0.42	0.51	
80%	0.97	0.78	0.77	0.82	0.72	0.80	
90%	1.27	1.17	1.07	1.16	1.14	1.14	
95%	1.52	1.36	1.38	1.41	1.41	1.44	
99%	1.93	1.88	1.87	1.90	1.82	2.11	
Case 4.							
1%	-4.06	-3.75	-2.79	-4.28	-3.62	-2.79	
5%	-2.52	-2.40	-2.15	-2.91	-2.49	-2.08	
10%	-1.82	-1.79	-1.64	-2.09	-1.81	-1.53	
20%	-0.99	-1.12	-0.99	-1.24	-1.14	-0.93	
30%	-0.53	-0.65	-0.63	-0.72	-0.75	-0.62	
40%	-0.17	-0.34	-0.30	-0.41	-0.39	-0.27	
50%	0.13	-0.04	-0.05	-0.09	-0.09	0.0	
60%	0.36	0.26	0.16	0.20	0.20	0.21	
70%	0.58	0.45	0.49	0.45	0.46	0.49	
80%	0.82	0.75	0.71	0.70	0.72	0.76	
90%	1.17	1.11	1.10	1.01	1.07	1.17	
95%	1.44	1.41	1.48	1.35	1.43	1.53	
99%	1.90	2.04	1.93	1.93	2.09	2.10	

Table 4.2. *Percentile of the distribution of* $t_{sw} = (\bar{y}_{sw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{sw}}}$, and $t_{ps} = (\bar{y}_{ps} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{psln}}}$ *for various cases and allocations*

Percentiles of	t_{sw}			t_{ps}			Z
	Sample allocation			Sample Allocation			
Case 1.	prop	propn	propd	prop	propn	propd	
1%	-1.04	4.12	8.57	-2.06	-3.21	-5.83	-2.33
5%	-0.74	4.29	8.84	-1.62	-2.05	-2.71	-1.65
10%	-0.59	4.40	8.97	-0.96	-1.56	-1.88	-1.28
20%	-0.32	4.54	9.14	-0.65	-0.95	-1.12	-0.84
30%	-0.18	4.64	9.28	-0.32	-0.61	-0.72	-0.52
40%	-0.03	4.72	9.41	-0.07	-0.28	-0.35	-0.25
50%	0.10	4.81	9.56	0.12	-0.01	-0.09	0.0
60%	0.23	4.90	9.69	0.20	0.24	0.16	0.25
70%	0.36	5.00	9.88	0.48	0.48	0.39	0.52
80%	0.57	5.13	10.05	0.81	0.79	0.68	0.84
90%	0.77	5.32	10.42	1.22	1.23	1.07	1.28
95%	0.95	5.47	10.71	1.59	1.46	1.43	1.65
99%	1.21	5.77	11.17	2.16	2.02	2.17	2.33
Case 2.							
1%	-0.85	1.43	4.63	-2.74	-2.43	-3.04	
5%	-0.51	1.61	4.80	-1.74	-1.74	-1.83	
10%	-0.34	1.72	4.92	-1.32	-1.27	-1.30	
20%	-0.15	1.86	5.09	-0.80	-0.80	-0.84	
30%	-0.01	1.96	5.21	-0.47	-0.48	-0.54	
40%	0.08	2.04	5.31	-0.23	-0.21	-0.29	
50%	0.20	2.12	5.40	0.06	0.06	0.05	
60%	0.30	2.20	5.49	0.28	0.31	0.23	
70%	0.41	2.28	5.61	0.56	0.57	0.50	
80%	0.51	2.36	5.72	0.83	0.81	0.84	
90%	0.69	2.49	5.92	1.25	1.23	1.35	
95%	0.84	2.61	6.08	1.64	1.61	1.82	
99%	1.02	2.83	6.37	2.14	2.32	2.54	

Table 4.2. Continued

Percentiles of	t_{sw}			t_{ps}			Z
	Sample allocation			Sample Allocation			
Case 3.	prop	propn	propd	prop	propn	propd	
1%	-1.63	2.00	3.20	-3.03	-2.60	-3.39	
5%	-1.13	2.21	3.40	-2.04	-1.76	-2.03	
10%	-0.79	2.31	3.47	-1.57	-1.32	-1.58	
20%	-0.53	2.43	3.59	-1.04	-0.87	-1.03	
30%	-0.34	2.52	3.68	-0.66	-0.53	-0.61	
40%	-0.14	2.61	3.74	-0.36	-0.27	-0.29	
50%	0.0	2.68	3.80	-0.08	0.0	-0.02	
60%	0.14	2.75	3.87	0.17	0.24	0.21	
70%	0.28	2.83	3.93	0.44	0.49	0.47	
80%	0.43	2.92	4.01	0.73	0.76	0.75	
90%	0.66	3.04	4.13	1.13	1.14	1.16	
95%	0.85	3.13	4.27	1.53	1.49	1.52	
99%	1.08	3.35	4.40	2.20	2.04	1.95	
Case 4.							
1%	-3.84	-0.20	1.40	-4.15	-3.93	-4.58	
5%	-2.49	0.56	1.77	-2.70	-2.39	-2.79	
10%	-1.84	0.87	1.96	-1.97	-1.80	-2.22	
20%	-1.10	1.22	2.16	-1.26	-1.15	-1.48	
30%	-0.63	1.39	2.30	-0.73	-0.76	-0.89	
40%	-0.30	1.56	2.43	-0.39	-0.38	-0.49	
50%	-0.04	1.69	2.54	-0.08	-0.10	-0.16	
60%	0.17	1.84	2.64	0.17	0.22	0.17	
70%	0.41	1.97	2.76	0.43	0.50	0.42	
80%	0.69	2.14	2.88	0.74	0.76	0.66	
90%	0.97	2.34	3.06	1.12	1.10	1.04	
95%	1.20	2.54	3.20	1.51	1.44	1.32	
99%	1.59	2.87	3.42	2.00	2.07	1.81	

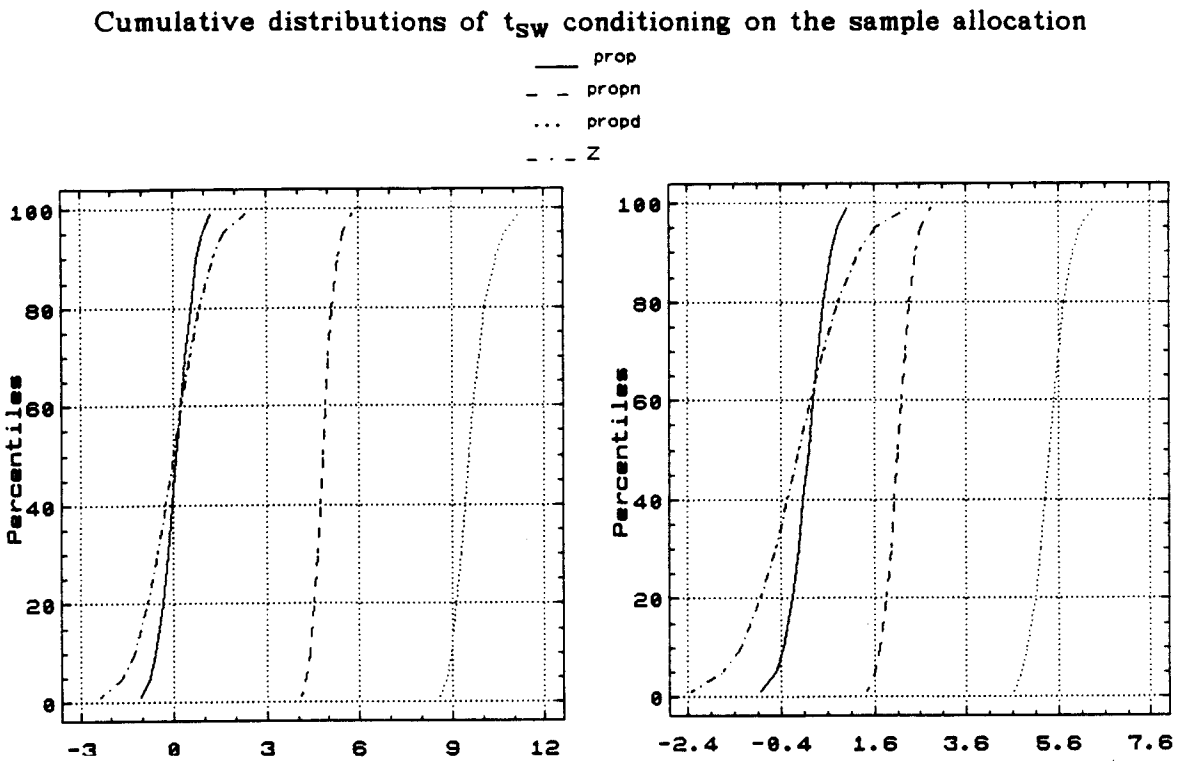

Figure 4.1 the distributions of t_{sw} for case 1 show that the bias is upward under propn and propd.

Figure 4.2. the distributions of t_{sw} for case 2 show that the bias is upward under propn and propd.

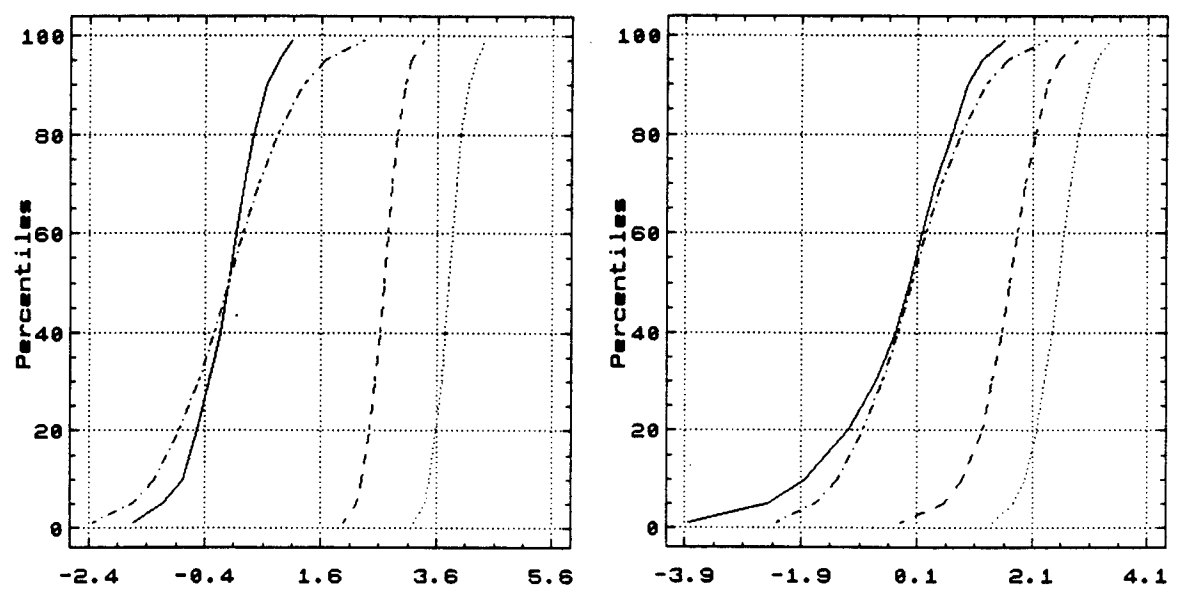

Figure 4.3 the distributions of t_{sw} for case 3 show that the bias is upward under propn and propd.

Figure 4.4. the distributions of t_{sw} for case 4 show that the bias is upward under propn and propd.

5. DISTRIBUTION OF POST STRATIFIED MEAN

In section 2.1 the distribution of $\bar{y}_{ps}$ was introduced and the conditional variance depending upon the sampling distribution was given in (2.2). Moreover the unconditional variance, which depends upon the unconditional distribution determined by all possible samples of fixed size n, is given in (2.3). Holt and Smith (1979) and Rao (1985) argued that the inferences should be made conditional on the achieved sample configuration.

The distribution of the post stratified mean using conditional inference for different sample sizes is introduced in the next section, averaging over all possible sample configurations. In the third section the distribution of the post stratified mean conditioned on the sample configuration is studied under three different configurations. In section four, comparison of post stratified mean and self-weighting mean is introduced.

5.1 The Distribution of $\bar{y}_{ps}$

The computer simulation mentioned in section 4.1 was also used to explore the distribution of

$$t_{ps} = (\bar{y}_{ps} - \bar{Y}) / \sqrt{ s^2_{\bar{y}_{ps|n}} } \qquad (5.1)$$

where $s^2_{\bar{y}_{ps|n}}$ is the post stratified sample variance given by

$$s^2_{\bar{y}_{ps|n}} = \sum_h W_h (1 - f_h) s_h^2 / n_h \qquad (5.2)$$

Table 4.1 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{ps} for the four cases under study which are given in Table 2.1. From Table 4.1 it is evident that t_{ps} has a distribution similar to the Z distribution even when relatively small sample sizes are employed for all cases.

5.2 The Distribution of $\bar{y}_{ps}$ Conditioned on the Sample Configuration

In this section the simulation study carried out in section 4.2 is employed to explore the distribution of $\bar{y}_{ps}$ conditioning on a particular sample allocation.

Table 4.2 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of the distributions of t_{ps} for the four cases under study and under the three allocations. The percentiles indicate that t_{ps} has approximately a Z distribution in cases 2 and 3 under the three allocations, whereas in cases 1 and 4 percentiles indicate that t_{ps} has some bias in the lower tail under propn and propd allocations. This bias might be due to the absence of the study variable in some sample units. For example, in case 4 there are several farms containing no wheat. However, Table 5.2 shows that the relative bias with respect to the population mean is very small and could be neglected.

Figures 5.1 - 5.4 emphasize the above argument and clearly show that t_{ps} has approximately Z distribution.

5.3 Comparison of Post Stratified and Self-Weighting Means

After sampling, if inferences are to be made conditional on the achieved sample configuration , n then it is natural to compare the performance of alternative estimators also conditional on n. Holt and Smith (1979) made comparisons between the conditional MSE of $\bar{y}_{sw}$ and the conditional variance of $\bar{y}_{ps}$; the difference in (2.19) was either positive or negative depending on the sample configuration , n. These were then compared empirically for a variety of real and artificial populations assuming that stratum sizes, means and variances were known .

In the next section the same comparison will be presented between the two estimators using "optimum" boundaries for stratification, and stratum means and variances estimated from the sample.

On the other hand, if the inferences are to be made unconditional the variance of $\bar{y}_{sw}$ that is defined by

$$V(\bar{y}) = (1 - f) S^2 / n$$

could be compared with the unconditional variance of $\bar{y}_{ps}$ defined in (2.3). This gives

$$V(\bar{y}_{sw}) - V(\bar{y}_{ps}) = \frac{(1 - f)}{n}\left\{ S^2 - \sum_h W_h S_h^2 \right\} - \frac{1}{n^2}\sum_h (1 - W_h) S_h^2 \qquad (5.3)$$

It is possible for (5.3) to be either positive or negative depending on the sample allocation relative to the post strata variances. As is true for the above comparison using conditional inferences, it is also true here that no estimator is uniformly better than the other.

5.4. An Empirical Comparison of $\bar{y}_{sw}$ and $\bar{y}_{ps}$

Equation (2.19) shows that neither $\bar{y}_{sw}$ nor $\bar{y}_{ps}$ is necessarily uniformly better in terms of conditional MSE. This section reports the results of comparisons made between the two estimators for a variety of real and artificial populations.

For any population with a known auxiliary variable, stratum boundaries will be determined by cum $\sqrt[3]{f(x)}$ approximation rule, and stratum means and variances will be estimated from the sample. For any given sample allocation n, calculate the sample variance $s^2_{\bar{y}_{sw}}$ of $\bar{y}_{sw}$ and the sample variance $s^2_{\bar{y}_{ps}}$ of $\bar{y}_{ps}$ and then calculate the design effect, K

$$K = s^2_{\bar{y}_{sw}} / s^2_{\bar{y}_{psln}} \qquad (5.4)$$

The computer simulation designed in section 4.1 is used to explore the distributions of K .

Table 5.1 contains the 1st, 5th, 10th, 90th, 95th, and 99th percentiles of K for a variety of populations are given in Table 2.1. In each population, the proportion of the total variance of the auxiliary variable accounted by the within stratum component , defined by

$$R = \frac{\sum_h N_h S_h^2}{\sum_h N_h S_h^2 + \sum_h N_h (\bar{X}_h - \bar{X})^2} \qquad (5.5)$$

This would be a measure of the potential value of stratification, i.e. where R is small, the post stratification is efficient.

Table 5.1 shows the potential impact of post stratification. For example for cases 1 and 2, percentiles indicate that the ratio of sample variance of $\bar{y}_{sw}$ to the sample variance of $\bar{y}_{ps}$ is greater than one. This suggests that $\bar{y}_{ps}$ is much better than $\bar{y}_{sw}$. The conclusion to be drawn from these two cases is that the gains of post stratification are great at no cost in precision. For case 3, which is an agricultural survey example, percentiles indicate that post stratified estimator is better at almost no cost where the corn production and the stratification variable , farm size , are highly correlated. However, in case 4 (a wheat production example) the gains from post stratification are greater in the upper tail where 10 per cent of the samples drawn led to a gain of at least 65 per cent, and five per cent of the samples led to gains in excess of 77 per cent . The cost of stratification at the lower tail, where 5 per cent of the samples K is less than 0.79, and 10 per cent of the samples, K is less than 0.91. This indicates that $\bar{y}_{sw}$ is slightly better than $\bar{y}_{ps}$, and this loss in the lower tail might be due to the fact that some farms contain no wheat.

Comparisons may be made of the percentiles of the distributions of t_{sw} and t_{ps} in Table 4.1 and the percentiles of the distribution of t_{sw} and t_{ps} conditioned on the sample allocations for the same variety of populations in Table 4.2 . From Table 4.1, percentiles indicate that t_{ps} has a distribution similar to the Z distribution even with relatively small sample size for all the cases. However, this is not the case for the distribution of t_{sw}. While the distribution of the latter is similar to the Z distribution in the two cases 1 and 2, it is not similar to Z in the other two cases 3 and 4 where sample sizes are relatively small . Table 4.2

contains percentiles of the distribution of t_{sw} and t_{ps} for the cases under study, and prop, propn and propd allocations respectively. The percentiles indicate that t_{sw} and t_{ps} have, in some cases, approximately the same distribution while in the other cases these distributions differed under the prop. For example, in case 4 t_{sw} and t_{ps} have almost the same distribution, but in case 2 the percentiles of t_{sw} have a shorter range. However, t_{sw} and t_{ps} would have the same distribution for large samples under this allocation . On the other hand, under the propn and propd allocations, percentiles indicate that t_{sw} has only positive percentiles, suggesting that $\bar{y}_{sw}$ is biased upward.

Table 5.2 contains the percentiles of the distribution of $K = s^2_{\bar{y}_{sw}} / s^2_{\bar{y}_{psln}}$ for the three different allocations. The percentiles indicate that $\bar{y}_{ps}$ is more efficient than $\bar{y}_{sw}$ for all cases and for all allocations, but the gains vary from one allocation to another. For example, in all the cases post stratification led to gains with a moderate range for the prop, but when the allocation altered from prop the gains encompassed a greater range. This result demonstrates that the self-weighting mean, $\bar{y}_{sw}$, is poor in unbalanced samples. On the other hand, the post stratified mean, $\bar{y}_{ps}$, appeared to balance out the conditional bias.

Comparing the graphs in Figures 4.1-4.4 with those in Figures 5.1 - 5.4 also emphasize the above results and show the conditional bias of $\bar{y}_{sw}$. On the other hand, the graphs show that $\bar{y}_{ps}$ is conditionally unbiased. The conditional bias of $\bar{y}_{sw}$ was evaluated relative to the population mean for the four cases under study and under the three different allocations and reported in Table 5.3 where in all cases the relative bias was small under prop,

and numerous under propn and propd allocations. For example, in case 3 the relative bias of $\bar{y}_{sw}$ was 0.4, 70.9, and 132 per cent of the population mean for prop, propn and propd respectively.

At this juncture, it is important to emphasize that another computer simulation was used to explore the distributions of K, t_{sw} and t_{ps} using the unconditional variances for the same populations. For any sample allocation, stratum sample variance and means were estimated, and K, t_{ps} were calculated as above by replacing the conditional sample variance $s^2_{\bar{y}_{ps|n}}$ with the unconditional variance $s^2_{\bar{y}_{ps}}$ that was defined in (2.3) for replacing S_h^2 by s_h^2.

The percentiles indicated that the post stratified mean was more efficient than the self-weighting mean in all cases except case 4 where post stratification provided considerable gains in the upper tail. However, the design effect, K , decreased in the lower tail, suggesting that $\bar{y}_{sw}$ is slightly better in this area. These results compare favorably with those discussed above using the conditional inferences.

Collectively, the comparisons indicate that post stratification is an efficient approach when the boundaries are obtained based on a proper stratification auxiliary variable, and it is more efficient than the SRS in reducing the MSE.

Furthermore, there is strong evidence that the post stratified estimator is a robust estimator against poorly distributed samples. However, in situations with heavily unequal post stratum variances it is possible for the design effect, K , to decrease. On the other hand , empirical investigations suggested that the self-weighting mean is very poor when the samples are

unbalanced.

Finally, assuming that the condition of the Central Limit Theorem holds true, the confidence intervals for any particular sample allocation based on the post stratified mean contain the population mean at the appropriate confidence level, but the confidence interval based on the self-weighting mean may not.

Table 5.1. *Percentiles of the distribution of K, the ratio of* $s^2_{\bar{y}_{sw}} / s^2_{\bar{y}_{psln}}$ *for various cases having proportion R of total variance within strata*

Case	No. of Strata	R	Sample Size	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	100	1.58	2.51	2.79	5.67	6.25	7.52
			200	1.51	2.60	2.88	4.96	5.32	6.03
			300	2.13	2.77	3.03	4.74	5.06	5.48
2.	3	.18	100	1.93	3.31	3.88	7.35	7.89	9.07
			200	3.09	4.00	4.39	6.82	7.17	7.77
			300	3.70	4.27	4.61	6.55	6.79	7.29
3.	4	.157	100	1.36	1.88	2.20	7.44	8.73	12.18
			200	.85	1.32	1.75	5.45	6.28	8.03
			500	.77	1.30	1.71	4.25	4.69	5.64
4.	4	.16	100	.55	.79	.91	2.85	3.69	5.98
			200	.46	.68	.83	1.98	2.40	3.39
			500	.53	.77	.89	1.65	1.77	2.08

Table 5.2. *Percentiles of the distribution of K, the ratio of* $s^2_{\bar{y}_{sw}} / s^2_{\bar{y}_{psln}}$ *for various cases and allocations*

Case	No. of Strata	R	Sample Allocation	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	prop	2.63	3.04	3.28	3.62	3.89	4.14
			propn	3.19	3.79	4.20	11.40	13.26	19.53
			propd	0.97	1.28	1.52	8.00	11.73	26.78
2.	3	.18	prop	4.18	4.71	5.00	7.65	8.39	9.80
			propn	6.67	7.34	7.90	12.59	13.45	15.87
			propd	3.91	4.50	4.85	10.13	11.76	14.75
3.	4	.157	prop	1.52	1.90	2.09	5.92	6.87	8.59
			propn	6.80	8.41	4.48	25.53	28.50	34.69
			propd	5.32	7.31	9.21	36.62	41.50	57.87
4.	4	.16	prop	1.02	1.05	1.07	1.89	2.54	4.31
			propn	1.62	2.00	2.32	9.88	11.61	15.61
			propd	0.87	1.36	1.78	17.21	21.34	31.11

Table 5.3. *Relative Bias of the estimates under study for various cases and allocations*

Case	Allocation	$\frac{\lvert \bar{y}_{sw}-\bar{Y}\rvert}{\bar{Y}}$%	$\frac{\lvert \bar{y}_{rsw}-\bar{Y}\rvert}{\bar{Y}}$%	$\frac{\lvert \bar{y}_{ps}-\bar{Y}\rvert}{\bar{Y}}$%	$\frac{\lvert \bar{y}_{prc}-\bar{Y}\rvert}{\bar{Y}}$%	$\frac{\lvert \bar{y}_{prs}-\bar{Y}\rvert}{\bar{Y}}$%
1.	prop	2.00	0.10	0.10	0.10	0.20
	propn	129.00	5.30	0.20	0.10	0.50
	propd	263.00	6.90	0.40	0.50	0.20
2.	prop	2.00	0.10	0.20	0.0	0.0
	propn	29.60	0.70	0.20	0.0	0.0
	propd	73.70	1.40	0.10	0.0	0.10
3.	prop	0.40	0.20	0.04	0.12	0.0
	propn	70.90	3.50	0.0	0.14	0.10
	propd	132.00	4.60	0.0	0.0	0.04
4.	prop	2.00	0.20	0.20	0.20	0.90
	propn	4.60	2.70	0.30	0.20	0.10
	propd	114.90	3.60	0.10	0.20	0.10

Cumulative distributions of t_{ps} conditioning on the sample allocation

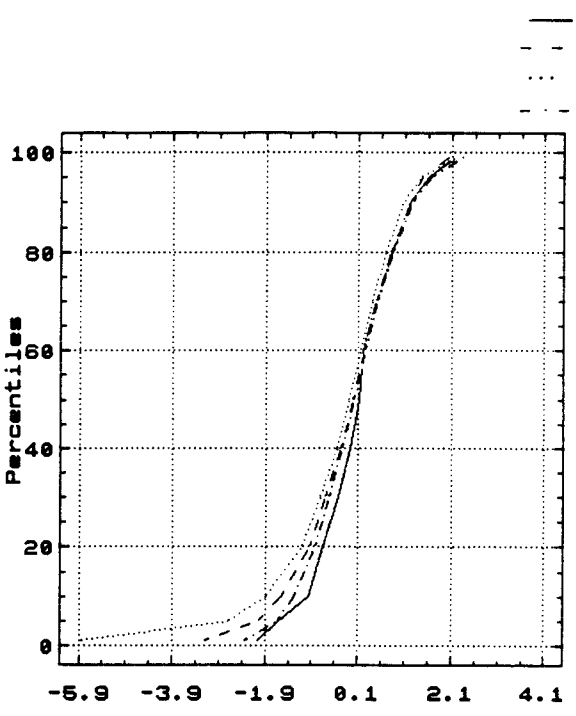

Figure 5.1 the distributions of t_{ps} for case 1 show that there is small bias under propn and propd.

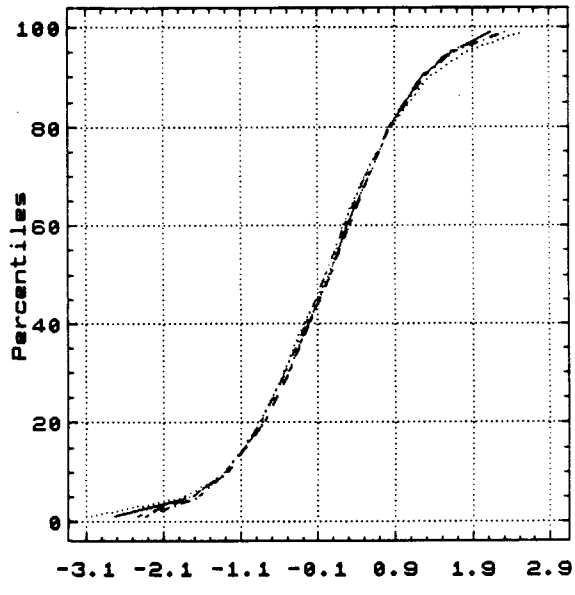

Figure 5.2. the distributions of t_{ps} for case 2 show that there is no bias under propn and propd.

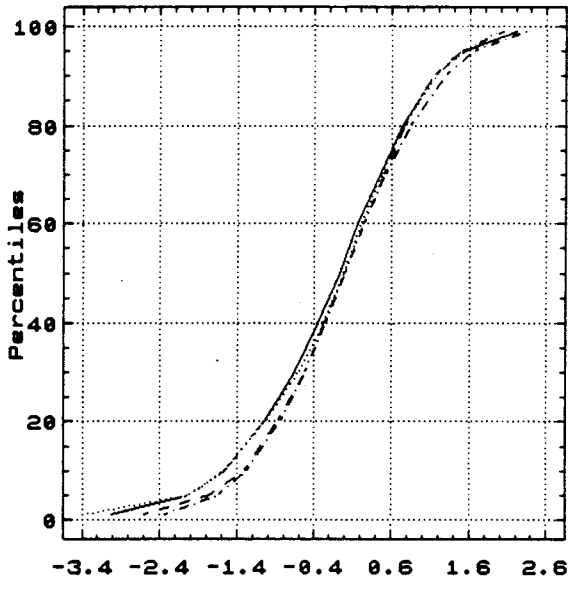

Figure 5.3 the distributions of t_{ps} for case 3 show that there is no bias under propn and propd.

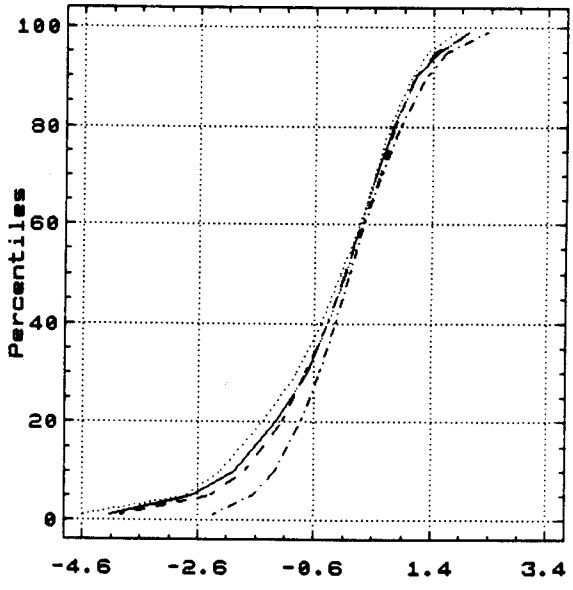

Figure 5.4. the distributions of t_{ps} for case 4 show that the is small bias under propn and propd.

6. DISTRIBUTION OF SELF-WEIGHTED RATIO ESTIMATOR

With an auxiliary variable that is highly correlated with the study variable, ratio estimates for the population mean or total improve the estimation. However, Royall and Cumberland (1981) pointed out that ratio estimates and the estimated variances can be conditionally biased unless the sample is balanced with respect to the auxiliary variable.

In the next section the distribution of the ratio estimate and its conditional and unconditional bias will be introduced.

In the third section a simulation study is made to explore the distribution of the self-weighting ratio estimate averaging over all possible sample allocations.

In the fourth section, the distribution of self-weighting ratio estimate is explored conditioning on a particular sample allocation.

In the fifth section, comparison of the post stratified mean and the self-weighting ratio is introduced from the analytical and empirical points of view for the populations under study.

6.1 Ratio Estimate and the Bias

After sampling, if inferences are to be made conditional on the sample allocation, n, then in order to compare the performance of the self-weighting ratio estimate, $\bar{y}_{rsw}$ its MSE should be also conditional. $\bar{y}_{rsw}$ can be written as

$$\bar{y}_{rsw} = \bar{X}\ \bar{y}_{sw} / \bar{x}_{sw} = \bar{X}\ \hat{R} \qquad (6.1)$$

where $\bar{y}_{sw}$ is defined in (2.10), $\hat{R}$ is the estimate of the ratio estimator $R = \bar{Y}/\bar{X}$, and conditional on n, the leading term in the conditional bias when it is expanded in a Taylor's series is obtained as follow

$$\hat{R} = \bar{y}_{sw} / \bar{x}_{sw}$$

$$\doteq (\bar{Y} / \bar{X}) + (\bar{y}_{sw} - \bar{Y})\left[\frac{\partial \hat{R}}{\partial \bar{y}_{sw}}\right]_{\bar{Y},\bar{X}} + (\bar{x}_{sw} - \bar{X})\left[\frac{\partial \hat{R}}{\partial \bar{x}_{sw}}\right]_{\bar{Y},\bar{X}}$$

$$+ \frac{1}{2}(\bar{y}_{sw} - \bar{Y})^2\left[\frac{\partial^2 \hat{R}}{\partial \bar{y}^2_{sw}}\right]_{\bar{Y},\bar{X}} + \frac{1}{2}(\bar{x}_{sw} - \bar{X})^2\left[\frac{\partial^2 \hat{R}}{\partial \bar{x}^2_{sw}}\right]_{\bar{Y},\bar{X}}$$

$$+ (\bar{y}_{sw} - \bar{Y})(\bar{x}_{sw} - \bar{X})\left[\frac{\partial^2 \hat{R}}{\partial \bar{y}_{sw}\,\partial \bar{x}_{sw}}\right]_{\bar{Y},\bar{X}}$$

The conditional expectation is

$$E(\hat{R} \mid n) \doteq R - \frac{1}{\bar{X}}\left[B_{\bar{y}} - RB_{\bar{x}}\right] + \frac{1}{\bar{X}^2}\sum_h w_h^2 \frac{(1 - f_h)}{n_h}\left[R^2 S_{xh}^2 - 2R\rho_h S_{yh}S_{xh}\right]$$

$$- \frac{B_{\bar{x}}}{\bar{X}^2}\left[B_{\bar{y}} - R\,B_{\bar{x}}\right]$$

where $B_{\bar{y}}$ and $B_{\bar{x}}$ are the conditional bias of $\bar{y}$ and $\bar{x}$ respectively given by

$$B_{\bar{y}} = \sum_h (w_h - W_h) \bar{Y}_h$$

$$B_{\bar{x}} = \sum_h (w_h - W_h) \bar{X}_h$$

Thus the conditional bias of $\bar{y}_{rsw}$ is given by

$$E(\bar{y}_{rsw} \mid n) - \bar{Y} \doteq \left[RB_{\bar{x}} - B_{\bar{y}}\right]\left[1 + \frac{B_{\bar{x}}}{\bar{X}} \right] + B_C \qquad (6.2)$$

where B_C is the leading term of the bias corresponding to that obtained by Cochran (1977) on page 161, that is defined by

$$B_C = \frac{1}{\bar{X}} \sum_h w_h^2 \frac{(1 - f_h)}{n_h} \left[R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh}\right] \qquad (6.3)$$

This bias is obtained from the second order approximation of Taylor's expansion. However, the conditional bias would be only the first term in the first part of equation (6.2). That is

$$\doteq \left[RB_{\bar{x}} - B_{\bar{y}}\right] \qquad (6.3)$$

when the first order approximation of Taylor's series is used.

A simulation study was made to explore the leading element in the conditional bias and the results are presented in Table 6.1. Empirical investigation suggested that the leading term of the conditional bias is the simple form in (6.3) that derived from the first order approximation.

Thus the conditional MSE of $\bar{y}_{rsw}$ using equation (6.3) is given by

$$MSE(\bar{y}_{rsw}|n) = \sum_h w_h^2 \frac{(1 - f_h)}{n_h} \left(S_{yh}^2 + R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh}\right) + \left(RB_{\bar{x}} - B_{\bar{y}}\right)^2 \qquad (6.4)$$

Cochran (1979) gave two alternative formulae for the sample estimate of the variance of $\bar{y}_{rsw}$. One form is

$$s^2_{\bar{y}_r} = \frac{(1 - f)}{n} (s_y^2 + \hat{R} s_x^2 - 2 \hat{R} s_{yx}) \qquad (6.5)$$

when $\bar{X}$ is not known, and the alternative form is

$$s^2_{\bar{y}_{ar}} = s^2_{\bar{y}_r} \frac{\bar{X}^2}{\bar{x}^2} \qquad (6.6)$$

where s_y^2, s_x^2 and s_{yx} are the usual sample variances and sample covariance of y and x respectively. In the next section the distribution of $\bar{y}_{rsw}$ will be explored and the question as to whether $s^2_{\bar{y}_r}$ in (6.5) is preferable to $s^2_{\bar{y}_{ar}}$ in (6.6) when $\bar{X}$ is known will be answered.

6.2. The Distribution of $\bar{y}_{rsw}$

The simulation was also used to explore the distributions of

$$t_{ar} = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{ar}}}, \text{ and } t_r = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_r}}$$

The entire process was performed as it was explained in section 4.1. for the

same cases.

Table 6.2 contains both the percentiles of t_{ar} and t_r for the cases under study, and the percentiles of Z.

Table 4.2 shows that the distributions for t_{ar} and t_r are very similar, and approximate to a Z distribution for relatively large samples. However, percentiles indicate that t_r is closer to the Z distribution for some cases, while t_{ar} is closer for others. For example, in case 1 the percentiles indicate that t_r is closer to Z than percentiles of t_{ar} for small sample sizes (≤ 100), but t_{ar} is closer to Z for relatively large samples (≥ 200). In case 2, the percentiles indicate the opposite, that is, t_{ar} is closer to Z for small sample sizes (≤ 100).

The conclusion to be drawn from this simulation is that there is little difference between the sample variance adjusted by $\bar{X}^2/\bar{x}^2$ as in (6.6) or the usual sample variance as defined in (6.5), for computing confidence intervals for any particular sample. This issue will be addressed in the next section conditioning on a particular sample allocation.

6.3. The Distribution of $\bar{y}_{rsw}$ Conditioned on the Sample Configuration

The computer simulation described above was also used to explore the distributions of

$$t_{ar} = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{ s^2_{\bar{y}_{ar}} }, \quad \text{and} \quad t_r = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{ s^2_{\bar{y}_r} }$$

conditioning on the sample allocation. The entire process was performed as it was explained in section 4.2. for the same cases.

Table 6.3 contains percentiles of the distribution of t_{ar} and t_r for the cases under study as well as prop, propn and propd allocations, respectively. The percentiles show that t_{ar} and t_r have approximately the same distribution in all cases under the prop. However, under the propn and propd allocations, the percentiles indicate that t_{ar} and t_r have only negative percentiles, thus illustrating the fact that $\bar{y}_{rsw}$ is biased downward for both distributions. However, their performance altered from one case to another, and in general $\bar{y}_{rsw}$ with the adjusted sample variance as given by (6.6) performed better in the real examples illustrated by cases 3 and 4 where data were obtained from agricultural surveys. In contrast, $\bar{y}_{rsw}$ with the usual sample variance as defined by (6.5) performed better for the artificial cases 1 and 2 where data were generated from linear models.

Figure 6.1- 6.4 clearly show the conditional bias of $\bar{y}_{rsw}$ in particular, when the sample allocation altered from the proportional allocation, and show that the bias is downward. Comparing the graphs in Figures 6.1(a)-6.4(a) with those in Figure 6.1(b) - 6.4(b), shows that $\bar{y}_{rsw}$ with adjusted sample variance is conditionally consistent while this is not the case with the usual sample variance for the real populations.

6.4. Comparison of Post Stratified and Self-Weighting Ratio Means

The conditional MSE of $\bar{y}_{rsw}$ as defined in (6.2) can be compared with the variance of $\bar{y}_{ps}$ defined in (2.2) as follows:

$$MSE\{(\bar{y}_{rsw} | n)\} - V\{(\bar{y}_{ps} | n)\} = \sum_h (w_h^2 - W_h^2) \frac{(1 - f_h)}{n_h} S_{yh}^2$$

$$+ \sum_h w_h^2 \frac{(1 - f_h)}{n_h} \left[R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh} \right]$$

$$+ \left[\sum_h (W_h - w_h)(\bar{Y}_h - R \bar{X}_h) \right]^2 \quad (6.7)$$

It is possible for equation (6.7) to be positive or negative depending on the sample allocation n, in which the difference is zero if the sample units are proportionally allocated. It cannot be said that one estimator is uniformly better than another.

An empirical comparison was carried out to examine the efficiency of the post stratified mean $\bar{y}_{ps}$ against the self-weighting ratio estimator $\bar{y}_{rsw}$. In this section we will report the results of the simulation study that was used to explore the distribution of the design effect K, where

$$K = s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{ps|n}}.$$

That is, the ratio of the sample variance of $\bar{y}_{rsw}$ to the sample variance of $\bar{y}_{ps}$.

Table 6.4 contains the 1st, 5th, 10th, 90th, 95th, 99th percentiles of K for the four cases under study. The percentiles indicate that the self-weighting ratio estimate is uniformly better in cases 1 and 2, where the design effect K is less than one. This is because the data is generated from linear models. In cases 3 and 4, where the data are

obtained from an agricultural survey, percentiles indicate that the post stratified mean is more efficient than the ratio estimator in the upper tail. Moreover, the post stratified mean led to considerable gains for relatively small samples. For example, in case 4 ten percent of the samples drawn led to a gain of at least five percent for post stratification, and five percent of the samples led to gains of eleven percent, and one percent of the samples to gains of at least twenty one percent. The gains doubled for relatively small samples. However, the ratio estimate is slightly better in the lower tail.

Table 6.5 contains the percentiles of the distribution of $K = s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{psln}}$ for the three different allocations. The percentiles indicate that $\bar{y}_{ps}$ is better than $\bar{y}_{rsw}$ for the real cases for all allocations, particularly in the upper tail. However, the gains vary from one allocation to another. For example, in cases 3 and 4 post stratification led to gains with a moderate range for the prop, but when the allocation altered from prop these gains encompassed a greater range. However, the percentiles indicate that $\bar{y}_{rsw}$ represents the best estimator for the artificial cases 1 and 2 for all allocations. This result shows that the self-weighting ratio $\bar{y}_{rsw}$, is very poor (under-estimates the mean) in unbalanced samples. On the other hand, the post stratified mean $\bar{y}_{ps}$, balanced out the conditional bias when it was used in sample survey situations.

The conditional bias of $\bar{y}_{rsw}$ was evaluated relative to the population mean for the four cases under study and under the three different allocations. The results are presented in Table 5.3 where in all cases the relative bias was small under prop, and large under propn and propd allocations. For example, in case 4 the relative bias of $\bar{y}_{rsw}$ was 0.2,

2.7 and 3.6 per cent of the population mean for prop, propn and propd, respectively. By comparison the relative bias of $\bar{y}_{ps}$, was 0.2, 0.3 and 0.1, respectively.

From the above results, it can be concluded that $\bar{y}_{ps}$ is conditionally unbiased for the population mean under any allocation. On the other hand, $\bar{y}_{rsw}$ is in general a biased estimator for the population mean. This bias is downward i.e. $\bar{y}_{rsw}$ is under-estimating the mean. Thus it is a poor estimator when the real cases are employed, as illustrated by case 3 and 4. Further, a comparison of $\bar{y}_{rsw}$ and $\bar{y}_{ps}$ shows that neither is necessarily better than the other in terms of the MSE, but empirical investigations carried out on a variety of real and artificial cases suggest that $\bar{y}_{ps}$ can be much better except when data is generated from linear models.

Finally, a comparison of t_{ar} and t_r shows that $\bar{y}_{rsw}$ with adjusted sample variance as defined in (6.6) is preferable to that with the usual sample variance, when data are obtained from an agricultural survey illustrated by cases 3 and 4 . However, percentiles in Table (6.3) and the graphs in Figures 6.1(b) - 6.2(b) show that $\bar{y}_{rsw}$ with the usual sample variance in (6.5) is preferable, when data are generated from linear models illustrated by cases 1 and 2.

On the other hand, if the inferences are to be made unconditional the variance of $\bar{y}_{rsw}$ that is defined in (7.6) could be compared with the unconditional variance of $\bar{y}_{ps}$ defined in (2.3).
That gives

$$MSE(\bar{y}_{rsw}) - V(\bar{y}_{ps}) = \frac{(1-f)}{n}\left[\left(S_y^2 - \sum_h W_h S_{yh}^2\right) + R^2 S_x^2 - 2R\rho S_y S_x\right]$$

$$- \frac{1}{n^2}\sum_h (1 - W_h) S_{yh}^2 \qquad (5.6)$$

The difference in (5.6) may be either positive , negative or zero depending on the variation between strata. For example, when the sample units are proportionally allocated, and S_h^2 is constant for all the strata, the difference becomes zero. It cannot be said one estimator is uniformly better than another.

An empirical comparison was carried out to examine the efficiency of the post stratified mean $\bar{y}_{ps}$, employing unconditional inferences, against the self-weighting ratio estimator $\bar{y}_{rsw}$.

The distribution of the design effect K, where

$$K = s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{ps}}$$

as well as the distributions of t_{ar} and t_{ps} using the unconditional variances for the same populations were explored. Empirical investigations indicated that the results obtained when the unconditional inferences were similar to the results obtained using conditional inferences, but conditional inferences led to larger gains with the post stratification technique.

Table 6.1. *Bias of $\bar{y}_{rsw}$ conditioned on the sample allocation with respect to the population mean for various cases*

Case	Allocation	$E(\bar{y}_{rsw}) - \bar{Y}$	$RB_{\bar{x}} - B_{\bar{y}}$	B_C	$E(\bar{y}_{rsw}\|n) - \bar{Y}$	$\bar{Y}$
1.	prop	-0.001	-0.013	0.001	-0.010	1.077
	propn	-0.057	-0.139	0.001	0.059	1.077
	propd	-0.074	-0.289	0.001	0.549	1.077
2.	prop	-0.002	-0.002	0.0002	-0.002	3.140
	propn	-0.023	-0.030	0.0003	-0.021	3.140
	propd	-0.043	-0.075	0.0003	-0.018	3.140
3.	prop	36.79	12.22	-13.30	-1.17	29613
	propn	1046.22	1735.03	-43.99	561.67	29613
	propd	1350.00	3062.00	-59.26	-728.85	29613
4.	prop	-2.130	-0.22	-1.38	-1.60	1014
	propn	-28.100	-43.79	-5.53	-20.81	1014
	propd	-35.010	-76.91	-7.89	8.93	1014

Table 6.2. *Percentile of the distribution of* $t_{rsw} = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rsw}}}$, and $t_r = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_r}}$ *for various cases*

Percentiles of	t_{rsw}			t_r			Z
	Sample Size			Sample Size			
Case 1.	100	200	300	100	200	300	
1%	-3.02	-2.49	-2.56	-2.23	-2.19	-2.24	-2.33
5%	-1.85	-1.81	-1.77	-1.59	-1.67	-1.60	-1.65
10%	-1.47	-1.36	-1.39	-1.28	-1.26	-1.30	-1.28
20%	-0.99	-0.87	-0.86	-0.91	-0.81	-0.82	-0.84
30%	-0.63	-0.56	-0.53	-0.59	-0.54	-0.51	-0.52
40%	-0.37	-0.27	-0.27	-0.36	-0.27	-0.27	-0.25
50%	-0.12	0.0	0.01	-0.12	0.0	0.01	0.0
60%	0.16	0.22	0.27	0.15	0.22	0.28	0.25
70%	0.41	0.50	0.53	0.44	0.53	0.55	0.52
80%	0.73	0.82	0.78	0.78	0.84	0.81	0.84
90%	1.12	1.20	1.27	1.28	1.32	1.34	1.28
95%	1.41	1.65	1.68	1.67	1.93	1.84	1.65
99%	1.98	2.27	2.31	2.51	2.80	2.73	2.33
Case 2.							
1%	-2.46	-2.59	-2.73	-2.39	-2.40	-2.63	
5%	-1.86	-1.79	-1.71	-1.77	-1.72	-1.65	
10%	-1.42	-1.33	-1.34	-1.32	-1.26	-1.31	
20%	-1.01	-0.92	-0.92	-0.94	-0.90	-0.91	
30%	-0.57	-0.59	-0.58	-0.56	-0.59	-0.58	
40%	-0.31	-0.32	-0.28	-0.29	-0.31	-0.28	
50%	-0.01	-0.02	-0.05	-0.01	-0.02	-0.05	
60%	0.22	0.19	0.20	0.23	0.20	0.21	
70%	0.47	0.47	0.46	0.48	0.47	0.46	
80%	0.73	0.77	0.79	0.77	0.80	0.79	
90%	1.27	1.27	1.22	1.34	1.25	1.28	
95%	1.63	1.59	1.60	1.77	1.70	1.66	
99%	2.21	2.32	2.46	2.46	2.56	2.58	

Table 6.2. Continued

Percentiles of	t_{rsw}			t_r			Z
	Sample Size			Sample Size			
Case 3.	100	200	300	100	200	300	
1%	-2.61	-2.55	-2.54	-2.57	-2.61	-2.62	
5%	-1.96	-1.89	-1.93	-1.92	-1.89	-1.94	
10%	-1.49	-1.39	-1.44	-1.46	-1.41	-1.47	
20%	-0.93	-0.91	-0.89	-0.91	-0.89	-0.89	
30%	-0.52	-0.57	-0.55	-0.50	-0.59	-0.56	
40%	-0.26	-0.30	-0.29	-0.24	-0.31	-0.29	
50%	0.06	-0.04	-0.05	0.05	-0.04	-0.04	
60%	0.28	0.25	0.24	0.27	0.24	0.23	
70%	0.57	0.52	0.48	0.53	0.51	0.48	
80%	0.90	0.82	0.80	0.83	0.80	0.79	
90%	1.41	1.23	1.21	1.30	1.23	1.19	
95%	1.70	1.58	1.48	1.63	1.54	1.48	
99%	2.43	2.08	1.94	2.26	2.00	2.00	
Case 4.							
1%	-3.71	-3.14	-2.61	-3.71	-3.19	-2.64	
5%	-2.64	-2.20	-1.91	-2.59	-2.29	-2.03	
10%	-1.92	-1.69	-1.44	-1.88	-1.65	-1.49	
20%	-1.13	-1.06	-0.95	-1.07	-1.09	-0.96	
30%	-0.71	-0.68	-0.63	-0.68	-0.67	-0.63	
40%	-0.37	-0.34	-0.30	-0.35	-0.33	-0.30	
50%	-0.09	-0.05	-0.04	-0.08	-0.05	-0.04	
60%	0.18	0.21	0.23	0.17	0.20	0.23	
70%	0.43	0.46	0.49	0.41	0.45	0.49	
80%	0.72	0.73	0.78	0.66	0.73	0.78	
90%	1.10	1.12	1.17	1.02	1.12	1.19	
95%	1.51	1.49	1.54	1.38	1.50	1.57	
99%	2.02	2.08	2.02	1.95	2.06	2.02	

Table 6.3. *Percentile of the distribution of* $t_{rsw} = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rsw}}}$, and $t_r = (\bar{y}_{rsw} - \bar{Y}) / \sqrt{s^2_{\bar{y}_r}}$ *for various cases and allocations*

Percentiles of	t_{rsw}			t_r			Z
	Sample allocation			Sample Allocation			
Case 1.	prop	propn	propd	prop	propn	propd	
1%	-2.43	-11.76	-25.63	-2.15	-4.80	-6.54	-2.33
5%	-1.67	-10.39	-23.39	-1.48	-4.20	-5.96	-1.65
10%	-1.30	- 9.84	-22.47	-1.18	-3.96	-5.73	-1.28
20%	-0.81	- 8.99	-21.44	-0.79	-3.70	-5.42	-0.84
30%	-0.53	- 8.43	-20.31	-0.51	-3.47	-5.18	-0.52
40%	-0.32	- 7.95	-19.56	-0.31	-3.29	-5.03	-0.25
50%	-0.09	- 7.56	-18.89	-0.08	-3.12	-4.84	0.0
60%	0.11	- 7.16	-18.28	0.12	-2.95	-4.70	0.25
70%	0.31	- 6.74	-17.71	0.31	-2.80	-4.55	0.52
80%	0.51	- 6.28	-16.97	0.53	-2.52	-4.37	0.84
90%	0.79	- 5.71	-16.09	0.84	-2.39	-4.13	1.28
95%	1.00	- 5.20	-15.27	1.08	-2.20	-3.93	1.65
99%	1.33	- 4.49	-13.86	1.42	-1.94	-3.60	2.33
Case 2.							
1%	-2.76	-4.48	-7.80	-2.64	-3.24	-4.36	
5%	-1.73	-3.59	-7.03	-1.66	-2.70	-3.93	
10%	-1.38	-3.13	-6.59	-1.33	-2.37	-3.70	
20%	-1.01	-2.67	-5.99	-0.98	-2.01	-3.38	
30%	-0.67	-2.33	-5.56	-0.67	-1.76	-3.17	
40%	-0.42	-2.02	-5.16	-0.41	-1.54	-2.94	
50%	-0.21	-1.75	-4.82	-0.20	-1.34	-2.75	
60%	0.05	-1.53	-4.48	0.05	-1.17	-2.55	
70%	0.32	-1.26	-4.20	0.31	-0.96	-2.41	
80%	0.59	-0.90	-3.80	0.59	-0.70	-2.16	
90%	0.90	-0.53	-3.29	0.90	-0.41	-1.89	
95%	1.21	-0.14	-2.96	1.18	-0.11	-1.68	
99%	1.69	0.31	-2.26	1.68	0.24	-1.32	

Table 6.3. Continued

Percentiles of	t_{rsw}			t_r			Z
	Sample allocation			Sample Allocation			
Case 3.	prop	propn	propd	prop	propn	propd	
1%	-2.82	-2.03	-1.42	-2.75	-1.23	-0.63	
5%	-1.83	-1.24	-0.75	-1.84	-0.75	-0.34	
10%	-1.44	-0.90	-0.43	-1.45	-0.54	-0.19	
20%	-0.94	-0.38	-0.07	-0.93	-0.23	-0.03	
30%	-0.58	-0.05	0.18	-0.59	-0.03	0.08	
40%	-0.26	0.23	0.44	-0.26	0.14	0.19	
50%	0.0	0.53	0.61	0.0	0.32	0.28	
60%	0.25	0.72	0.84	0.24	0.44	0.36	
70%	0.53	0.97	1.06	0.52	0.59	0.47	
80%	0.84	1.29	1.30	0.85	0.79	0.58	
90%	1.29	1.70	1.64	1.29	1.04	0.72	
95%	1.67	2.06	2.01	1.64	1.25	0.89	
99%	2.31	2.85	2.51	2.28	1.69	1.10	
Case 4.							
1%	-3.58	-3.59	-3.04	-3.54	-2.22	-1.40	
5%	-2.37	-3.32	-1.91	-2.43	-1.45	-0.89	
10%	-1.83	-1.77	-1.48	-1.89	-1.09	-0.67	
20%	-1.15	-1.14	-0.95	-1.19	-0.69	-0.43	
30%	-0.69	-0.76	-0.64	-0.69	-0.47	-0.29	
40%	-0.38	-0.44	-0.42	-0.38	-0.27	-0.19	
50%	-0.11	-0.19	-0.20	-0.11	-0.11	-0.09	
60%	0.18	0.05	0.0	0.18	0.03	0.0	
70%	0.43	0.31	0.22	0.43	0.19	0.10	
80%	0.71	0.57	0.43	0.74	0.35	0.19	
90%	1.16	1.02	0.75	1.15	0.61	0.33	
95%	1.42	1.35	1.08	1.38	0.78	0.48	
99%	1.94	1.85	1.45	1.96	1.10	0.64	

Table 6.4. *Percentiles of the distribution of K, the ratio of* $s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{psln}}$ *for various cases having proportion R of total variance within strata*

Case	No. of Strata	R	Sample Size	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	100	0.02	0.03	0.04	0.12	0.14	0.17
			200	0.03	0.04	0.04	0.09	0.10	0.13
			300	0.04	0.04	0.05	0.09	0.09	0.11
2.	3	.18	100	0.01	0.01	0.01	0.02	0.02	0.03
			200	0.01	0.01	0.01	0.02	0.02	0.02
			300	0.01	0.01	0.01	0.02	0.02	0.02
3.	4	.157	100	0.16	0.31	0.41	1.25	1.46	1.91
			200	0.18	0.28	0.36	1.02	1.12	1.37
			500	0.19	0.31	0.39	0.85	0.90	1.03
4.	4	.16	100	0.42	0.52	0.62	1.39	1.56	2.27
			200	0.31	0.52	0.62	1.21	1.32	1.48
			500	0.39	0.59	0.68	1.05	1.11	1.21

Table 6.5. *Percentiles of the distribution of K, the ratio of* $s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{psln}}$ *for various cases and allocations*

Case	No. of Strata	R	Sample Allocation	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	prop	0.02	0.02	0.03	0.13	0.17	0.25
			propn	0.003	0.002	0.003	0.01	0.01	0.02
			propd	0.001	0.001	0.001	0.002	0.002	0.004
2.	3	.18	prop	0.01	0.01	0.01	0.02	0.03	0.04
			propn	0.01	0.01	0.01	0.02	0.02	0.02
			propd	0.001	0.001	0.001	0.002	0.003	0.004
3.	4	.157	prop	0.16	0.29	0.39	1.12	1.29	1.78
			propn	0.41	0.56	0.65	1.93	2.22	2.63
			propd	0.13	0.17	0.22	1.86	2.21	3.16
4.	4	.16	prop	0.43	0.51	0.60	1.33	1.48	1.92
			propn	0.50	0.57	0.63	1.83	2.16	2.78
			propd	0.13	0.17	0.22	1.86	2.21	3.16

Cumulative distributions of $\bar{y}_{rsw}$ conditioning on the sample allocation

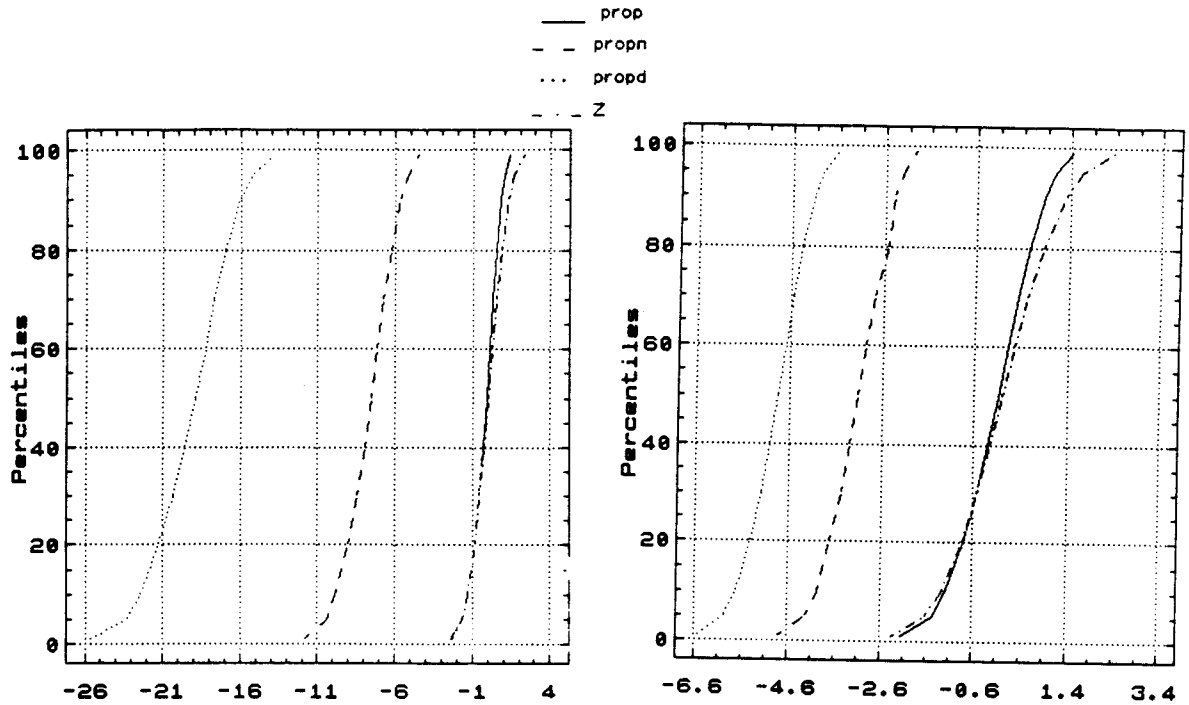

Figure 6.1(a) the distributions of t_{ar} for case 1 show that the bias is downward under propn and propd.

Figure 6.1.(b) the distributions of t_r for case 1 show that the bias is downward under propn and propd.

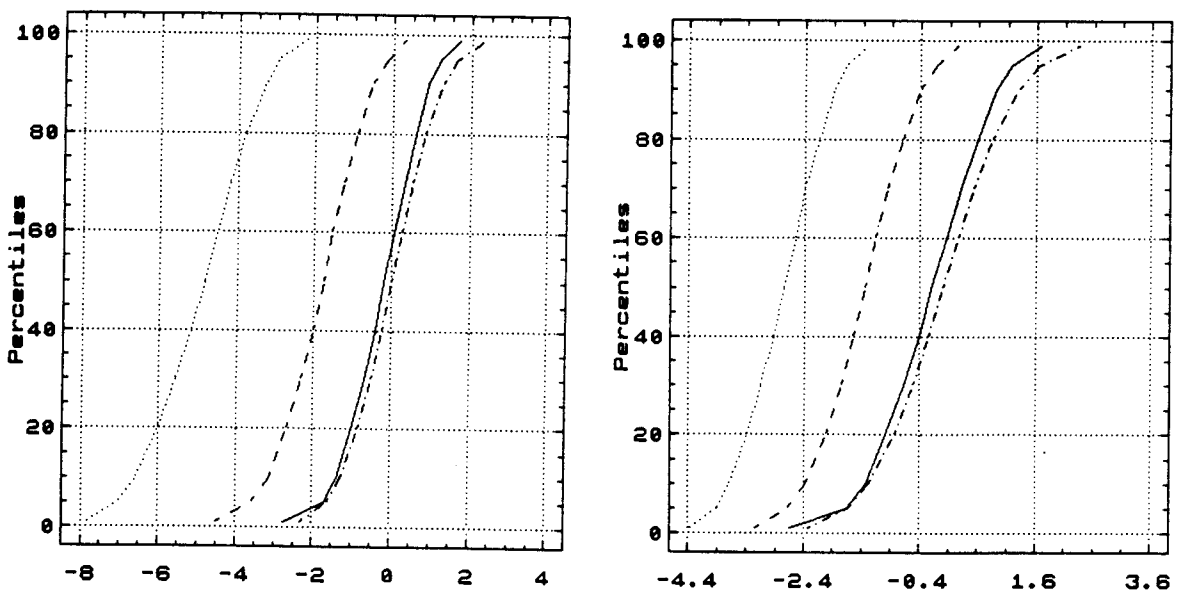

Figure 6.2(a) the distributions of t_{ar} for case 2 show that the bias is downward under propn and propd.

Figure 6.2.(b) the distributions of t_r for case 2 show that the bias is downward under propn and propd.

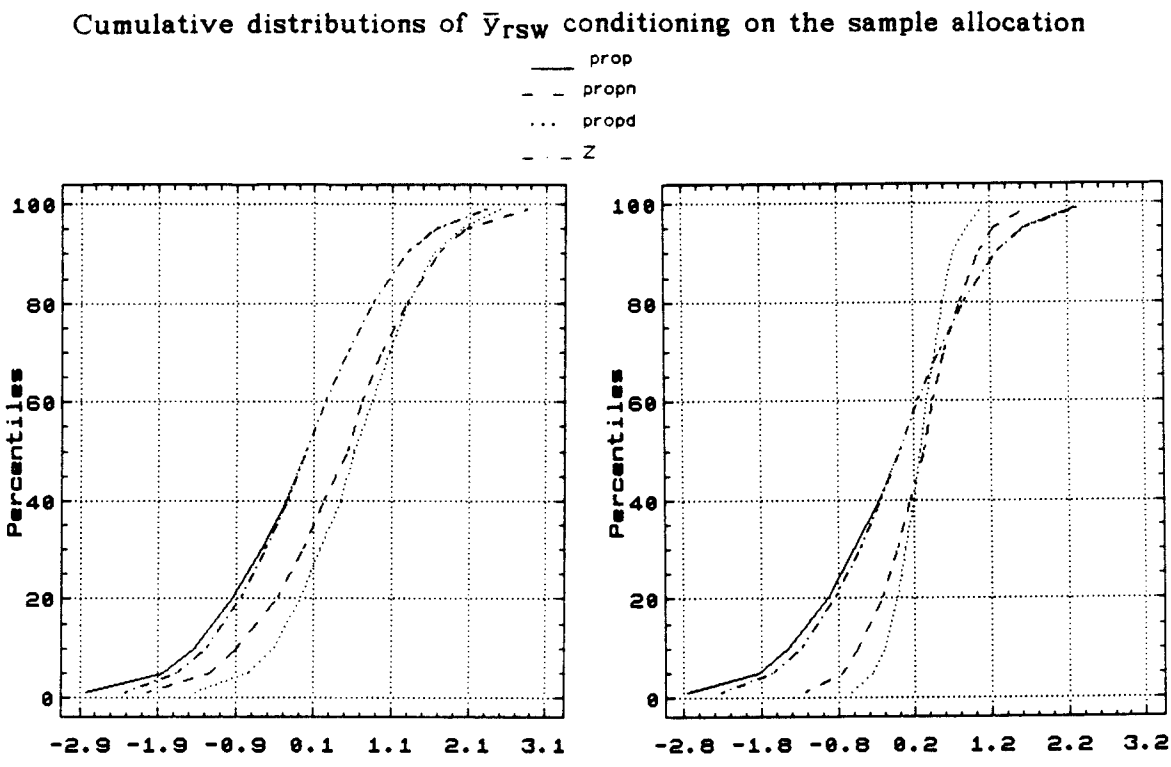

Figure 6.3(a) the distributions of t_{ar} for case 3 show that the bias is declined under propn and propd.

Figure 6.3.(b) the distributions of t_r for case 1 show that the bias is declined but it is skewed.

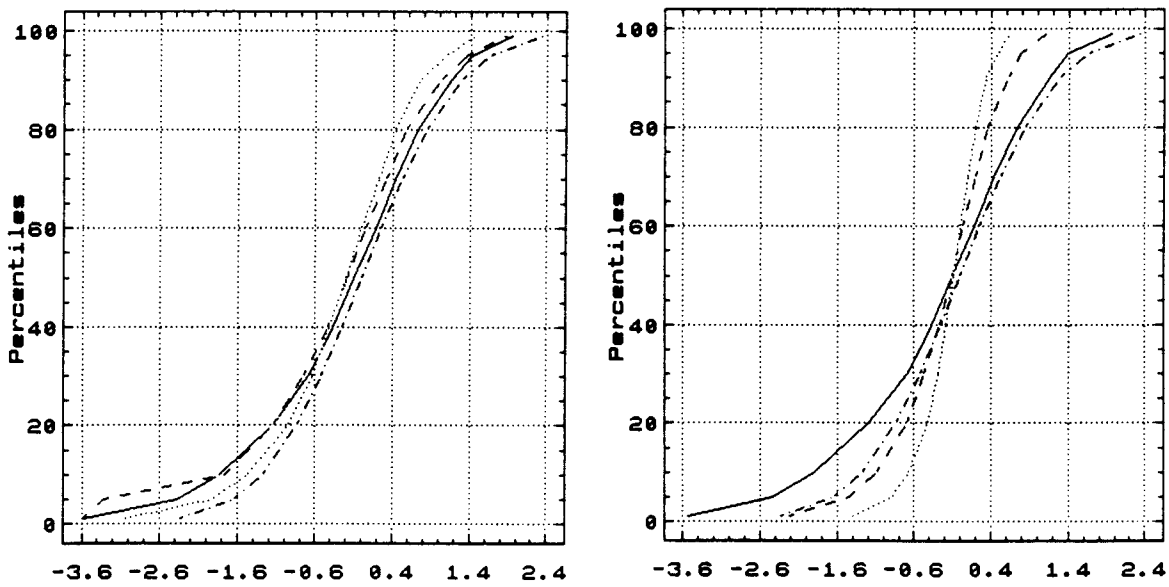

Figure 6.4(a) the distributions of t_{ar} for case 3 show that the bias is declined under propn and propd.

Figure 6.4.(b) the distributions of t_r for case 1 show that the bias is declined but it is skewed.

7. DISTRIBUTION OF POST STRATIFIED COMBINED RATIO ESTIMATOR

In chapter 6. it was argued that the self-weighting ratio estimate and its estimated variance can be conditionally biased unless the sample is balanced with respect to the auxiliary variable.

Robinson (1987) adjusted the conditional bias using only randomization theory with a regression model, but he mentioned that if the complete set of $X_1, \ldots, X_N$ is known, it would be appropriate to use post stratification as suggested by Fuller (1981) in which the analysis of Royall and Cumberland (1981) were discussed.

In this chapter the distribution of the post stratified combined ratio estimator will be introduced and compared with the self-weighting ratio estimator.

In the next section, the distribution of the post stratified combined ratio estimator conditioning on the sample allocation will be introduced and the sample variance obtained.

A computer simulation will be used in the third section to explore the distribution of the combined ratio estimator averaging over all possible sample allocations.

In the fourth section, the distribution of the combined ratio estimator conditioning on the sample allocation will be introduced employing a simulation study.

Analytical and empirical comparisons between post stratified combined and self-weighting ratio estimates are introduced in the fifth section.

7.1 Post Stratified Combined Ratio Estimator is Conditionally Unbiased

Let $\bar{y}_{prc}$ be the post stratified combined ratio estimator. In section (6.1) the conditional MSE of $\bar{y}_{rsw}$ was derived . In this section, the conditional variances of $\bar{y}_{prc}$ will be derived.

$\bar{y}_{prc}$ can be written as

$$\bar{y}_{prc} = (\bar{y}_{ps} / \bar{x}_{ps}) \bar{X}$$

It is approximately unbiased conditioned on the sample configuration.
The conditional variance is simply the usual variance of the STRS that is given by

$$V\{(\bar{y}_{prc} \mid n)\} \doteq V\{(\bar{y}_{ps} \mid n)\} + R^2 V\{(\bar{x}_{ps} \mid n)\} - 2 R \, COV\{(\bar{y}_{ps}, \bar{x}_{ps}) \mid n)$$

$$= \sum_h W_h^2 \left(\frac{1-f_h}{n_h}\right)\left[S_{yh}^2 + R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh} \right] \quad (7.1)$$

The sample variance can be obtained from (7.1) with S_{yh}^2, S_{xh}^2 and R estimated by s_{yh}^2, s_{xh}^2 and $\hat{R}$, respectively. This gives

$$s^2_{\bar{y}_{rcln}} = \sum_h W_h^2 \left(\frac{1-f_h}{n_h}\right)\left[s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2 \hat{R} s_{yxh} \right] \quad (7.2)$$

where s_{yxh} is the hth sample covariance.

The unconditional variances of $\bar{y}_{prc}$ can be derived by averaging equation (7.1) over all possible distributions of n
Employing equation (2.3) this gives

$$V(\bar{y}_{prc}) = \frac{1-f}{n}\sum_{h} W_h \left(S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh} S_{xh} \right)$$

$$+ \frac{1}{n^2}\sum_{h} (1 - W_h) \left(S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh} S_{xh} \right) \qquad (7.3)$$

The sample variances can be obtained as described previously for the conditional variance. That is

$$s^2_{\bar{y}_{rc}} = \frac{1-f}{n}\sum_{h} W_h \left(s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{yxh} \right)$$

$$+ \frac{1}{n^2}\sum_{h} (1 - W_h) \left(s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{yxh} \right) \qquad (7.4)$$

7.2 The Distribution of $\bar{y}_{prc}$

The distribution of $t_{rc} = (\bar{y}_{prc} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rcln}}}$ was explored by the empirical study described in section 4.1 and the results are reported in Table 7.1.

Table 7.1 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{rc} for the four cases under study, which are also given in Table 2.1.

Table 7.1 shows that t_{rc} has a distribution similar to the Z distribution in all cases with relatively large sample sizes. However, t_{rc} also has a Z distribution in the linear model cases illustrated by 1 and 2, even with relatively small sample sizes.

7.3 The Distribution of $\bar{y}_{prc}$ Conditioned on the Sample Configuration

We have argued that the self-weighting ratio estimator is conditionally biased, and the conditional bias is as defined in (6.3). In this section the computer simulation described in section 4.2 is employed to explore the distribution of $\bar{y}_{rc}$ conditioning on a particular sample allocation.

Table 7.2 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{rc} for the four cases under study and under three allocations.

Table 7.2 shows that t_{rc} had approximately the Z distribution under all allocations studied. Thus empirical investigation suggests that $\bar{y}_{rc}$ is conditionally unbiased. Figures 7.1- 7.4 clearly illustrate the above results.

7.4 Comparison of Post Stratified Combined and Self-Weighting Ratio Estimates

After sampling, the inferences are to be made conditional on the sample configuration, n. In this section we will compare the post stratified combined ratio estimator with the self-weighting ratio estimator from analytical and empirical points of view.

Comparing the MSE of $\bar{y}_{rsw}$ in (6.4) with the variance of $\bar{y}_{prc}$ in (7.1) gives

$$MSE\{(\bar{y}_{rsw}|\, n\,)\} - V\{(\bar{y}_{prc}|\, n\,)\} =$$

$$\sum_h \left(w_h^2 - W_h^2 \right) \frac{\left(1 - f_h \right)}{n_h} \left[S_{yh}^2 + R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh} \right]$$

$$+ \left[\sum_h \left(W_h - w_h \right) \left(\bar{Y}_h - R \bar{X}_h \right) \right]^2 \qquad (7.5)$$

The difference in (7.5) is either positive or negative depending on the sample allocation, n, relative to the post strata means and variances. The difference is zero if the sample units are proportionally allocated to the strata.

The ratio, $K = s^2_{\bar{y}_{ar}} / s^2_{\bar{y}_{rcln}}$, of the sample variance of $\bar{y}_{rsw}$ in (6.6) to the sample variance of $\bar{y}_{prc}$ in (7.2) was adjusted by the ratio $\bar{X}_h^2 / \bar{x}_h^2$ for each stratum and its distribution was explored. The results are presented in Table 7.3.

Table 7.3 contains the percentiles of the distribution of K for the populations under study. The percentiles show the potential impact of post stratified ratio estimation over the self-weighting ratio estimation with SRS. For example, in cases 1 and 2, percentiles indicate that the post stratified combined ratio estimator is uniformly better than the self-weighting ratio even with data generated from linear models for these two cases. When we turn to agricultural survey situations described by cases 3 and 4, the potential of the post stratification technique becomes clear in terms of its robust nature.

Table 7.4 contains the percentiles of the distribution of $K = s^2_{\bar{y}_{ar}} / s^2_{\bar{y}_{prcln}}$ for the three different allocations. The percentiles indicate that $\bar{y}_{prc}$ was better than $\bar{y}_{rsw}$ for the real cases for all allocations, particularly in the upper tail, although the gains did vary from one allocation to another. For example, in cases 3 and 4 post

stratification led to gains with a moderate range for the prop, but when the allocation altered from prop the gains increased. However, the percentiles indicate that $\bar{y}_{rsw}$ was better in the linear model cases with unbalanced samples.

From Table 5.3 the relative bias of $\bar{y}_{rsw}$ was small for balanced sample, prop, and increased for unbalanced samples. For example, in case 3 the relative bias of $\bar{y}_{rsw}$ was 0.2, 3.5 and 4.6 per cent of the population mean for prop, propn and propd respectively.
By comparison, the relative bias of $\bar{y}_{prc}$ was 0.12, 0.14 and zero respectively. This result demonstrates that the self-weighting ratio $\bar{y}_{rsw}$, is poor in unbalanced samples while, on the other hand, the post stratified ratio $\bar{y}_{prc}$ balanced out the conditional bias when it was used.

Comparing the percentiles in Table 6.5. and Table 7.4 shows that t_{prc} had approximately a Z distribution under all allocations, but t_{rsw} had a Z distribution only under the proportional allocation. This result suggests that the self-weighting ratio estimate should not be used to asses the population mean unless the sample is balanced.

Comparing the graphs of the Figures 6.1-6.4 with those of Figures 7.1-7.4 emphasize the above results.

On the other hand, if the inferences are to be made unconditional, the variance of $\bar{y}_{rsw}$ which corresponds to the usual formula

$$MSE(\bar{y}_{rsw}) = \frac{(1-f)}{n}(S_y^2 + R^2 S_x^2 - 2\rho S_x S_y) \qquad (7.6)$$

could be compared with the unconditional variance of $\bar{y}_{prc}$ defined in (7.3).

This gives

$$MSE(\bar{y}_{rsw}) - V(\bar{y}_{prc}) =$$

$$\frac{(1-f)}{n}\left[S_y^2 - \sum_h W_h S_{yh}^2 + R^2\left(S_x^2 - \sum_h W_h S_{xh}^2\right) - 2R\left(\rho S_y S_x - \sum_h W_h \rho_h S_{yh} S_{xh}\right)\right]$$

$$- \frac{1}{n^2}\sum_h (1 - W_h)\left(S_{yh}^2 + R^2 S_{xh}^2 - 2 R \rho_h S_{yh} S_{xh}\right) \qquad (7.7)$$

The difference in (7.7) is either positive or negative depending on the sample allocation, n , relative to the post strata variances. The difference is zero if the sample units are proportionally allocated to strata. As was the case for the above comparison using conditional inferences, it also is true here that no estimator is uniformly better than the other. Another computer simulation was used to explore the distributions of K, t_{rsw} and t_{rc} using the unconditional variances for the same populations. For any sample allocation, stratum sample variance and means are estimated and K, t_{rc} are calculated as above by replacing the conditional sample variance $s^2_{\bar{y}_{rc|n}}$ with the unconditional variance $s^2_{\bar{y}_{rc}}$ that is defined in (7.4) .

The percentiles also indicate that the post stratified combined ratio in general is better than the self-weighting ratio estimate in all cases. These results compare favorably with those discussed above using the conditional inferences.

Table 7.1. *Percentile of the distribution of* $t_{rc} = (\bar{y}_{prc} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rcln}}}$, and $t_{rs} = (\bar{y}_{prs} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rsln}}}$ *for various cases*

Percentiles of	t_{rc}			t_{rs}			Z
	Sample Size			Sample Size			
Case 1.	100	200	300	100	200	300	
1%	-2.71	-2.61	-2.74	-2.68	-2.58	-2.50	-2.33
5%	-1.76	-1.77	-1.72	-1.72	-1.70	-1.72	-1.65
10%	-1.35	-1.27	-1.33	-1.37	-1.27	-1.33	-1.28
20%	-0.89	-0.83	-0.84	-0.93	-0.85	-0.79	-0.84
30%	-0.58	-0.52	-0.53	-0.60	-0.54	-0.49	-0.52
40%	-0.32	-0.29	-0.25	-0.29	-0.24	-0.24	-0.25
50%	-0.02	0.01	0.0	-0.01	-0.05	0.01	0.0
60%	0.24	0.27	0.24	0.26	0.27	0.25	0.25
70%	0.53	0.50	0.53	0.51	0.51	0.52	0.52
80%	0.86	0.79	0.82	0.75	0.81	0.83	0.84
90%	1.24	1.28	1.30	1.22	1.20	1.22	1.28
95%	1.57	1.56	1.60	1.60	1.59	1.65	1.65
99%	2.23	2.05	2.12	2.09	2.10	2.14	2.33
Case 2.							
1%	-2.50	-2.52	-2.48	-2.45	-2.51	-2.50	
5%	-1.77	-1.74	-1.75	-1.78	-1.69	-1.76	
10%	-1.44	-1.35	-1.28	-1.39	-1.31	-1.31	
20%	-0.96	-0.91	-0.87	-0.98	-0.92	-0.87	
30%	-0.60	-0.59	-0.57	-0.59	-0.56	-0.54	
40%	-0.29	-0.28	-0.28	-0.26	-0.26	-0.25	
50%	-0.04	-0.02	-0.04	0.01	0.01	0.0	
60%	0.23	0.20	0.24	0.25	0.23	0.23	
70%	0.49	0.45	0.51	0.49	0.50	0.50	
80%	0.78	0.74	0.78	0.82	0.87	0.82	
90%	1.21	1.20	1.21	1.23	1.17	1.21	
95%	1.56	1.56	1.61	1.61	1.60	1.61	
99%	2.46	2.35	2.32	2.47	2.30	2.38	

Table 7.1. Continued

Percentiles of	t_{rc}			t_{rs}			Z
	Sample Size			Sample Size			
Case 3.	100	200	300	100	200	300	
1%	-2.55	-2.87	-2.60	-2.91	-3.13	-2.65	
5%	-2.01	-1.86	-1.89	-2.04	-1.96	-1.91	
10%	-1.56	-1.38	-1.42	-1.58	-1.39	-1.44	
20%	-0.98	-0.91	-0.86	-1.00	-0.94	-0.87	
30%	-0.57	-0.56	-0.54	-0.57	-0.58	-0.54	
40%	-0.28	-0.28	-0.26	-0.31	-0.28	-0.27	
50%	0.0	-0.03	0.0	-0.04	-0.05	-0.02	
60%	0.27	0.21	0.24	0.24	0.22	0.23	
70%	0.53	0.48	0.52	0.52	0.50	0.52	
80%	0.84	0.82	0.80	0.87	0.82	0.82	
90%	1.42	1.34	1.23	1.45	1.34	1.27	
95%	1.78	1.63	1.51	1.82	1.64	1.55	
99%	2.35	2.18	1.91	2.38	2.29	2.14	
Case 4.							
1%	-4.62	-3.30	-2.64	-4.11	-3.39	-2.75	
5%	-2.63	-2.37	-1.99	-2.76	-2.40	-2.03	
10%	-1.96	-1.70	-1.47	-2.01	-1.74	-1.49	
20%	-1.16	-1.08	-0.96	-1.19	-1.15	-0.96	
30%	-0.72	-0.75	-0.62	-0.77	-0.77	-0.67	
40%	-0.38	-0.36	-0.32	-0.44	-0.39	-0.34	
50%	-0.05	-0.09	-0.01	-0.10	-0.11	-0.04	
60%	0.21	0.18	0.23	0.18	0.16	0.22	
70%	0.45	0.46	0.49	0.43	0.43	0.46	
80%	0.74	0.75	0.78	0.73	0.70	0.75	
90%	1.10	1.16	1.18	1.09	1.11	1.17	
95%	1.49	1.54	1.52	1.46	1.49	1.50	
99%	2.00	2.14	2.06	1.95	2.13	2.05	

Table 7.2. *Percentile of the distribution of*

$t_{rc} = (\bar{y}_{prc} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{prcin}}}$, and $t_{rs} = (\bar{y}_{prs} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{prsin}}}$

for various cases and allocations

Percentiles of	t_{rc}			t_{rs}			Z
	Sample allocation			Sample Allocation			
Case 1.	prop	propn	propd	prop	propn	propd	
1%	-2.93	- 3.26-	3.78	-2.88	-3.04	-3.88	-2.33
5%	-2.06	- 2.24-	2.35	-2.06	-2.14	-2.29	-1.65
10%	-1.53	- 1.73-	1.77	-1.50	-1.63	-1.64	-1.28
20%	-0.92	- 1.11-	1.04	-0.94	-1.03	-0.94	-0.84
30%	-0.56	- 0.69-	0.59	-0.56	-0.62	-0.53	-0.52
40%	-0.28	- 0.33-	0.25	-0.28	-0.27	-0.22	-0.25
50%	0.02	0.02	0.04	0.01	0.02	0.06	0.0
60%	0.30	0.32	0.28	0.29	0.29	0.30	0.25
70%	0.58	0.63	0.53	0.55	0.55	0.50	0.52
80%	0.86	0.95	0.89	0.83	0.80	0.80	0.84
90%	1.31	1.41	1.38	1.20	1.15	1.08	1.28
95%	1.57	1.77	1.79	1.53	1.43	1.34	1.65
99%	2.14	2.44	2.66	1.97	1.82	2.02	2.33
Case 2.							
1%	-2.80	-2.56	-2.98	-2.73	-2.61	-2.85	
5%	-1.80	-1.83	-1.87	-1.76	-1.80	-1.85	
10%	-1.42	-1.40	-1.39	-1.43	-1.39	-1.42	
20%	-1.03	-0.99	-0.90	-0.95	-0.98	-0.89	
30%	-0.65	-0.66	-0.58	-0.63	-0.62	-0.59	
40%	-0.36	-0.36	-0.30	-0.32	-0.32	-0.29	
50%	-0.11	-0.11	0.0	-0.05	-0.07	0.02	
60%	0.19	0.20	0.25	0.21	0.20	0.27	
70%	0.54	0.52	0.56	0.57	0.54	0.54	
80%	0.87	0.90	0.89	0.91	0.88	0.89	
90%	1.32	1.35	1.34	1.29	1.35	1.25	
95%	1.68	1.84	1.92	1.68	1.72	1.53	
99%	2.42	2.48	2.61	2.35	2.27	2.18	

Table 7.2. Continued

Percentiles of	t_{rc}			t_{rs}			Z
	Sample allocation			Sample Allocation			
Case 3.	prop	propn	propd	prop	propn	propd	
1%	-3.04	-2.64	-2.97	-3.13	-2.65	-3.16	
5%	-1.97	-1.78	-2.01	-2.02	-1.79	-2.03	
10%	-1.45	-1.38	-1.49	-1.50	-1.41	-1.54	
20%	-0.91	-0.87	-0.99	-0.96	-0.93	-0.99	
30%	-0.55	-0.53	-0.59	-0.57	-0.54	-0.60	
40%	-0.25	-0.23	-0.24	-0.26	-0.25	-0.27	
50%	0.0	0.0	0.0	-0.02	0.0	-0.02	
60%	0.24	0.25	0.24	0.24	0.23	0.24	
70%	0.50	0.56	0.49	0.49	0.56	0.51	
80%	0.86	0.90	0.80	0.87	0.90	0.80	
90%	1.39	1.24	1.30	1.44	1.30	1.26	
95%	1.81	1.65	1.73	1.95	1.69	1.71	
99%	2.71	2.18	2.43	2.71	2.15	2.40	
Case 4.							
1%	-3.81	-3.34	-4.11	-4.03	-3.37	-4.31	
5%	-2.49	-2.30	-2.67	-2.60	-2.31	-2.81	
10%	-1.95	-1.80	-2.14	-2.03	-1.85	-2.23	
20%	-1.14	-1.15	-1.43	-1.23	-1.18	-1.49	
30%	-0.70	-0.71	-0.90	-0.76	-0.73	-0.93	
40%	-0.36	-0.40	-0.46	-0.40	-0.41	-0.49	
50%	-0.10	-0.05	-0.10	-0.13	-0.06	-0.11	
60%	0.18	0.20	0.18	0.14	0.19	0.17	
70%	0.42	0.53	0.43	0.40	0.51	0.43	
80%	0.76	0.77	0.72	0.74	0.76	0.71	
90%	1.19	1.12	1.02	1.12	1.11	1.01	
95%	1.46	1.39	1.33	1.46	1.38	1.35	
99%	1.95	2.01	1.80	1.95	2.01	1.77	

Table 7.3. *Percentiles of the distribution of* $K = s^2_{\bar{y}_{ar}} / s^2_{\bar{y}_{prcln}}$ *for various cases having proportion R of total variance within strata*

Case	No. of Strata	R	Sample Size	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	100	1.08	1.24	1.36	2.33	2.54	2.86
			200	1.26	1.45	1.53	2.18	2.27	2.46
			300	1.41	1.53	1.60	2.10	2.16	2.30
2.	3	.18	100	0.88	1.04	1.12	1.81	1.93	2.14
			200	1.08	1.18	1.23	1.66	1.76	1.88
			300	1.16	1.23	1.27	1.61	1.67	1.77
3.	4	.157	100	0.56	0.69	0.77	1.77	2.16	2.98
			200	0.30	0.60	0.84	1.42	1.86	2.22
			500	0.44	0.60	0.70	1.29	1.41	1.62
4.	4	.16	100	0.54	0.67	0.74	1.41	1.62	2.24
			200	0.45	0.65	0.77	1.27	1.38	1.67
			500	0.61	0.77	0.83	1.13	1.17	1.27

Table 7.4. *Percentiles of the distribution of K, the ratio of* $s^2_{\bar{y}_{rsw}} / s^2_{\bar{y}_{rcln}}$ *for various cases and allocations*

Case	No. of Strata	R	Sample Allocation	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	prop	1.16	1.31	1.41	2.27	2.42	2.72
			propn	0.06	0.08	0.09	0.27	0.31	0.40
			propd	0.001	0.004	0.006	0.04	0.06	0.10
2.	3	.18	prop	0.96	1.07	1.12	1.78	1.90	2.14
			propn	0.47	0.54	0.60	1.05	1.15	1.29
			propd	0.04	0.06	0.07	0.23	0.26	0.35
3.	4	.157	prop	0.54	0.64	0.69	1.53	1.87	2.57
			propn	0.80	0.96	1.12	3.14	3.62	4.54
			propd	0.39	0.55	0.68	2.65	3.20	3.99
4.	4	.16	prop	0.67	0.74	0.79	1.31	1.43	2.18
			propn	0.56	0.64	0.70	1.95	2.26	2.81
			propd	0.14	0.20	0.24	1.82	2.14	2.82

Cumulative distributions of $\bar{y}_{prc}$ conditioning on the sample allocation

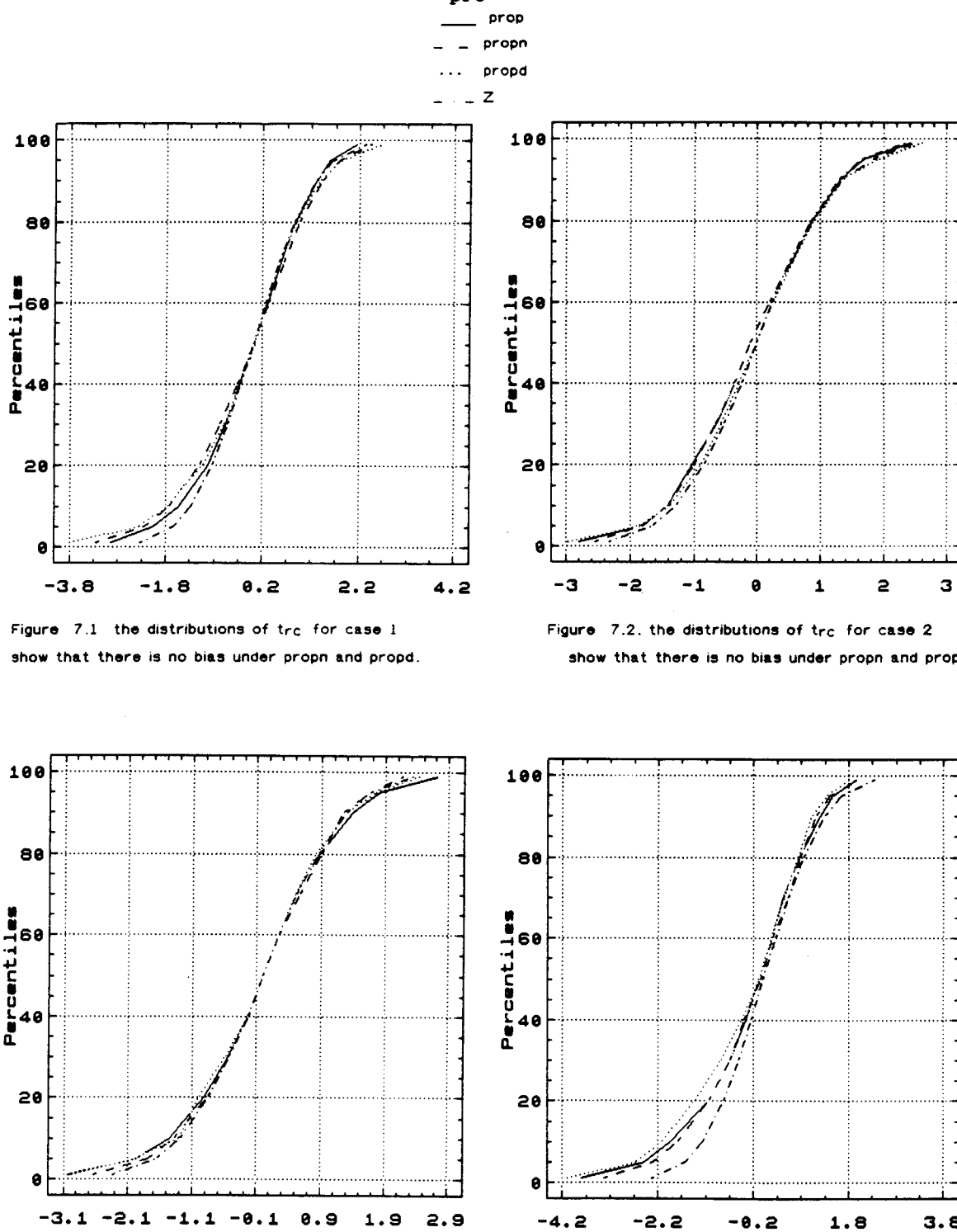

Figure 7.1 the distributions of t_{rc} for case 1 show that there is no bias under propn and propd.

Figure 7.2. the distributions of t_{rc} for case 2 show that there is no bias under propn and propd.

Figure 7.3 the distributions of t_{rc} for case 3 show that there is very small bias under propd.

Figure 7.4. the distributions of t_{rc} for case 4 show that there is small bias under propd.

8. DISTRIBUTION OF POST STRATIFIED SEPARATE RATIO ESTIMATOR

The separate ratio estimator is another good alternative to the self-weighting ratio estimate when the sample is unbalanced with respect to the auxiliary variable. Robinson (1987) suggested that if the complete set of $X_1, \ldots, X_N$ is known, it would be appropriate to use post stratification as suggested by Fuller (1981) in which the analysis of Royall and Cumberland (1981) were discussed.

In this chapter the distribution of the post stratified separate ratio estimator will be introduced and compared with the self-weighting ratio estimator.

In the next section, the distribution of the post stratified separate ratio estimator conditioning on the sample allocation will be introduced and sample variance obtained.

In the third section, a simulation study will be employed to explore the distribution of the separate ratio estimator averaging over all possible sample allocations.

In the fourth section, the distribution of the separate ratio estimator conditioning on the sample allocation will be introduced employing a simulation study.

In the fifth section, analytical and empirical comparisons between post stratified separate and self-weighting ratio estimates are introduced.

8.1 Post Stratified Separate Ratio Estimator is Conditionally Unbiased

Let $\bar{y}_{prs}$ be the post stratified separate ratio estimator that can be written as

$$\bar{y}_{prs} = \frac{1}{N} \sum_{h} \frac{\bar{y}_h}{\bar{x}_h} X_h$$

where X_h is the population stratum total. The conditional variance is simply the usual variance of the STRS that is given by

$$V\{(\bar{y}_{prs} \mid n)\} \doteq V\{(\bar{y}_{ps} \mid n)\} + R^2 V\{(\bar{x}_{ps} \mid n)\} - 2 R \, COV\{(\bar{y}_{ps}, \bar{x}_{ps}) \mid n)$$

Thus, conditional on the sample allocation the variance is given by

$$V\{(\bar{y}_{prs} \mid n)\} \doteq \sum_{h} W_h^2 \left(\frac{1 - f_h}{n_h}\right)\left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2 R_h \rho_h S_{yh} S_{xh}\right] \quad (8.1)$$

The sample variance can be obtained from (8.1) with S_{yh}^2, S_{xh}^2 and R estimated by s_{yh}^2, s_{xh}^2 and $\hat{R}$, respectively. This gives

$$s^2_{\bar{y}_{rsln}} = \sum_{h} W_h^2 \left(\frac{1-f_h}{n_h}\right)\left[s_{yh}^2 + \hat{R}_h^2 s_{xh}^2 - 2 \hat{R}_h s_{yxh} \right] \quad (8.2)$$

where s_{yxh} is the hth sample covariance.

The unconditional variances of $\bar{y}_{prs}$ can be derived by averaging equation (8.1) over all possible distributions of n

Employing equation (2.3) this gives

$$V(\bar{y}_{prs}) = \frac{1 - f}{n} \sum_{h} W_h \left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2 R_h \rho_h S_{yh} S_{xh} \right]$$

$$+ \frac{1}{n^2} \sum_{h} (1 - W_h) \left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2 R_h \rho_h S_{yh} S_{xh} \right] \quad (8.3)$$

The sample variance may be estimated by

$$s^2_{\bar{y}_{rs}} = \frac{1-f}{n} \sum_h W_h \left(s^2_{yh} + \hat{R}_h^2 s^2_{xh} - 2 \hat{R}_h s_{yxh} \right)$$

$$+ \frac{1}{n^2} \sum_h (1 - W_h) \left(s^2_{yh} + \hat{R}_h^2 s^2_{xh} - 2 \hat{R}_h s_{yxh} \right) \qquad (8.4)$$

8.2 The Distribution of $\bar{y}_{prs}$

The distribution of $t_{rs} = (\bar{y}_{prs} - \bar{Y}) / \sqrt{s^2_{\bar{y}_{rsln}}}$ was explored by the empirical study described in section 4.1 and the results are reported in Table 7.1.

Table 7.1 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{rs} for the four the populations under study .

From Table 7.1 percentiles of the distributions of t_{rs} and t_{rc} indicate similar distributions that approximately fit a Z distribution.

8.3 The Distribution of $\bar{y}_{prs}$ Conditioned on the Sample Configuration

In this section the computer simulation described in section 4.2 is employed to explore the distribution of $\bar{y}_{rs}$ conditioning on a particular sample allocation.

Table 7.2 contains the 1st, 5th, 10th, 20th, . . .,90th, 95th, 99th percentiles of t_{rs} for the four cases under study and under three allocations.

From Table 7.2 it is apparent that the percentiles indicate t_{rs} and

t_{rc} had an approximate Z distribution under all allocations studied. Thus empirical investigation suggests that $\bar{y}_{rs}$ is conditionally unbiased. Figures 7.1-7.4 and Figures 8.1- 8.4 illustrate those observations.

8.4 Comparison of Post Stratified Separate and Self-weighting Ratio Estimates

After sampling, the inferences are to be made conditional on the sample configuration, n . In this section we will compare post stratified separate ratio estimator with the self-weighting ratio estimator from both analytical and empirical points of view.

Comparing the MSE of $\bar{y}_{rsw}$ in (6.4) with the variance of $\bar{y}_{prc}$ in (8.1) gives

$$MSE\{(\bar{y}_{rsw} | n)\} - V\{(\bar{y}_{prs} | n)\} =$$

$$\sum_h (w_h^2 - W_h^2)\frac{(1-f_h)}{n_h}\left[S_{yh}^2 + (R^2 - R_h^2)S_{xh}^2 - 2(R-R_h)\rho_h S_{yh}S_{xh}\right] + \left[\sum_h (W_h - w_h)(\bar{Y}_h - R\bar{X}_h)\right]^2 \qquad (8.5)$$

The difference in (8.5) is either positive or negative depending on the sample allocation, n , relative to the post strata means and variances. The difference is zero if the sample units are proportionally allocated to the strata. The ratio, $K = s^2_{\bar{y}_{ar}} / s^2_{\bar{y}_{rsln}}$, of the sample variance of $\bar{y}_{rsw}$ in (6.6) to the sample variance of $\bar{y}_{prs}$ in (8.2) was adjusted by the ratio $\bar{X}_h^2 / \bar{x}_h^2$ for each stratum and its distribution was explored. The results are presented in Table 8.1.

Table 8.1 contains the percentiles of the distribution of K for the populations under study. The percentiles show the potential impact of post stratified ratio estimation over self-weighting ratio estimation with SRS. For example, in cases 1 and 2, percentiles indicate that the post stratified separate ratio estimator is uniformly better than the self-weighting ratio even with data generated from linear models for these two cases. When we turn to agricultural survey situations illustrated by cases 3 and 4, the potential of the post stratification technique becomes clear in terms of its robust nature.

Table 8.2 contains the percentiles of the distribution of $K = s^2_{\bar{y}ar} / s^2_{\bar{y}prsln}$ for the three different allocations. The percentiles indicate that $\bar{y}_{prs}$ was better than $\bar{y}_{rsw}$ for the real cases for all allocations, particularly in the upper tail, although the gains did vary from one allocation to another. For example, in cases 3 and 4 post stratification led to gains with a moderate range for the prop, but when the allocation altered from prop the gains increased. However, the percentiles indicate that $\bar{y}_{rsw}$ was better in the linear model cases with unbalanced samples.

From Table 5.3 the relative bias of $\bar{y}_{rsw}$ for case 3 was 0.2, 3.5 and 4.6 per cent of the population mean for prop, propn and propd respectively. By comparison the relative bias of $\bar{y}_{prs}$ was 0.0, 0.10 and 0.04 respectively. This result demonstrates that post stratified separate ratio estimator $\bar{y}_{prs}$, balanced out the conditional bias when it was used.

Comparing the percentiles in Table 6.5. and Table 7.4 shows that t_{prs} had approximately a Z distribution under all allocations, but t_{rsw}

had only Z distribution under the proportional allocation.

Comparing the graphs of the Figures 6.1-6.4 with those of Figures 8.1-8.4 emphasize the above results.

On the other hand, if the inferences are to be made unconditional the variance of $\bar{y}_{rsw}$ that is defined in (7.6) could be compared with the unconditional variance of $\bar{y}_{prc}$ defined in (8.3), to give

$$MSE(\bar{y}_{rsw}) - V(\bar{y}_{prs}) = \frac{(1-f)}{n}\Big\{S_y^2 - \sum_h W_h S_{yh}^2 + R^2 S_x^2 - \sum_h W_h R_h^2 S_{xh}^2 - 2\Big(R\rho S_y S_x - \sum_h W_h R_h \rho_h S_{yh} S_{xh}\Big)\Big\}$$

$$- \frac{1}{n^2}\sum_h (1 - W_h)\left(S_{yh}^2 + R_h^2 S_{xh}^2 - 2 R_h \rho_h S_{yh} S_{xh}\right) \qquad (8.6)$$

The difference in (8.6) may be either positive or negative depending on the sample allocation, n, relative to the post strata variances. The difference is zero if the sample units are proportionally allocated to the strata and R_h is constant from one stratum to another.

Another computer simulation was used to explore the distributions of K, t_{rsw} and t_{rs} using the unconditional variances for the same populations. The conditional sample variance $s^2_{\bar{y}_{rs|n}}$ was replaced by the unconditional variance $s^2_{\bar{y}_{rs}}$ that was defined in (8.4). In this instance, the same results were obtained when the unconditional inferences were used to compare the two estimates with the conditional ones. However, the gains were slightly larger when conditional inferences were applied with the post stratification.

The percentiles also indicate that the post stratified combined ratio, in general, is better than the self-weighting ratio estimate in all cases. These results compare favorably with those discussed above using the conditional inferences.

Table 8.1. *Percentiles of the distribution of* $K = s^2_{\bar{y}rsw} / s^2_{\bar{y}prsln}$ *for various cases having proportion* R *of total variance within strata*

Case	No. of Strata	R	Sample Size	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	100	0.65	0.81	0.90	1.78	1.92	2.28
			200	0.86	0.98	1.05	1.67	1.79	2.01
			300	0.98	1.05	1.11	1.61	1.69	1.86
2.	3	.18	100	0.80	0.94	1.02	1.73	1.83	2.15
			200	0.96	1.08	1.13	1.57	1.68	1.81
			300	1.06	1.13	1.17	1.52	1.57	1.67
3.	4	.157	100	0.58	0.69	0.78	1.93	2.45	3.64
			200	0.30	0.52	0.67	1.51	1.80	2.37
			500	0.44	0.63	0.72	1.33	1.48	1.68
4.	4	.16	100	0.54	0.60	0.75	1.51	1.75	2.52
			200	0.47	0.68	0.78	1.31	1.46	1.93
			500	0.64	0.79	0.84	1.13	1.19	1.32

Table 8.2. *Percentiles of the distribution of* K, *the ratio of* $s^2_{\bar{y}ar} / s^2_{\bar{y}prsln}$ *for various cases and allocations*

Case	No. of Strata	R	Sample Allocation	Percentiles of K 1%	5%	10%	90%	95%	99%
1.	3	.26	prop	0.81	0.93	1.00	1.54	1.63	1.84
			propn	0.02	0.04	0.04	0.15	0.18	0.22
			propd	0.001	0.002	0.002	0.02	0.03	0.06
2.	3	.18	prop	0.90	0.98	1.04	1.59	1.69	1.99
			propn	0.41	0.47	0.51	0.91	1.00	1.12
			propd	0.03	0.04	0.05	0.19	0.22	0.29
3.	4	.157	prop	0.54	0.64	0.70	1.77	2.20	3.28
			propn	0.80	0.97	1.12	3.17	3.66	4.57
			propd	0.39	0.56	0.68	2.63	3.17	4.04
4.	4	.16	prop	0.68	0.76	0.81	1.40	1.61	2.32
			propn	0.56	0.64	0.70	2.01	2.38	2.93
			propd	0.15	0.21	0.24	1.93	2.41	3.46

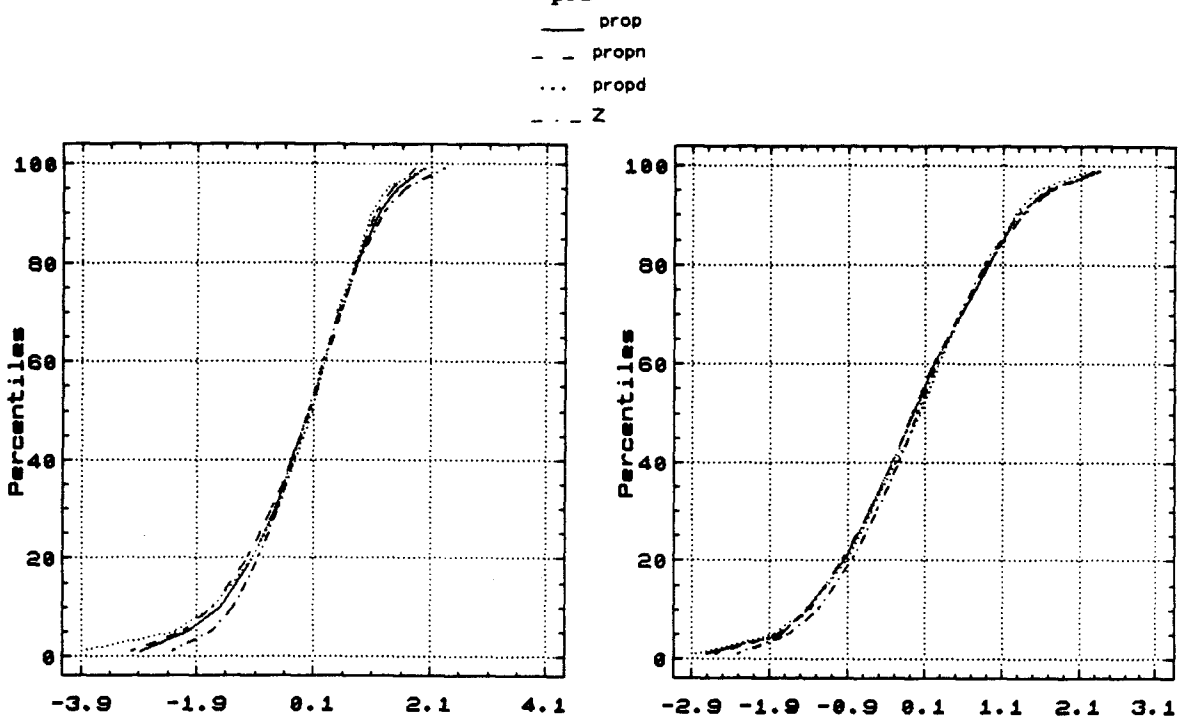

Figure 8.1 the distributions of t_{rs} for case 1 show that there is no bias under propn and propd.

Figure 8.2. the distributions of t_{rs} for case 2 show that there is no bias under propn and propd.

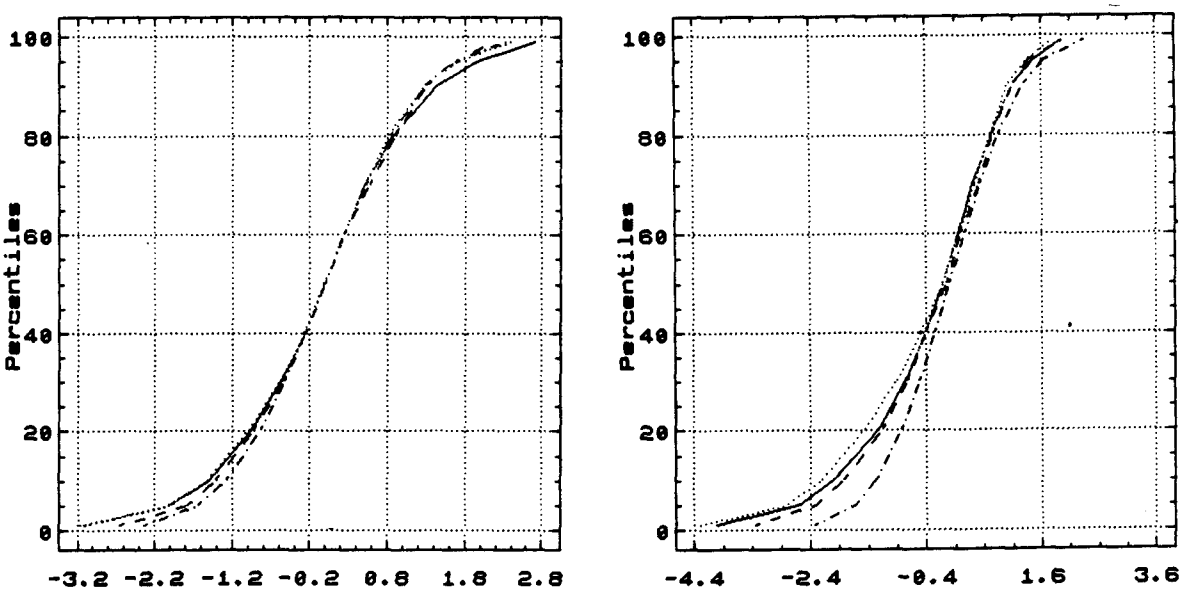

Figure 8.3 the distributions of t_{rs} for case 3 show that there is very small bias under propd.

Figure 8.4. the distributions of t_{rs} for case 4 show that there is small bias under propd.

9. CONCLUSIONS

When one is interested in estimating the population mean for a study variable that is related to a known auxiliary variable, the post stratification technique using stratum boundaries based on the auxiliary variable can provide estimates whose variances are less than the estimates with categorical or arbitrary boundaries.

If post stratification is efficient, post stratified estimates are better than the self-weighting estimates with SRS in reducing the conditional MSE. There is evidence that post stratification is a robust technique which provides balance against occasional poorly distributed samples.

In this thesis , we considered the question dealing with the use of a known auxiliary variable to post stratify the population of the study variable in order to reduce the conditional bias. In chapter 3 a method for constructing strata based on the auxiliary variable and Dalenius's equations was introduced. An approximation method was employed, based on Thomsen's (1976) rule to minimizing the variance under proportional allocation, that is, to use cum $\sqrt[3]{f(x)}$ rule as approximation to chose the best boundaries.

It was argued that the appropriate framework for the comparisons of post stratified and self-weighting estimates is the distribution conditional on the achieved sample configuration, n. The unconditional sampling distribution may be used at the design stage before the sample is drawn.

In chapter 4 the distribution of the self-weighting mean was obtained and its conditional bias was explored. The percentiles of the distributions

in Table 4.2 and the graphs in the Figures 4.1- 4.4 show that $\bar{y}_{sw}$ is conditionally biased upward.

In chapter 5 the distribution of the post stratified mean was obtained and the percentiles of the distributions in Table 4.2 and the graphs in the Figures 5.1-5.4 indicate that $\bar{y}_{ps}$ is conditionally unbiased for the population mean even with the unbalanced samples. Even though equation (2.13) show that $\bar{y}_{ps}$ is not uniformly better, from Table 5.1 , there was strong evidence that the post stratified mean is more efficient than the self-weighting mean in reducing the MSE.

In chapter 6 the distribution of the self-weighting was studied and a leading term for the conditional bias was derived, and then its conditional MSE was obtained. The percentiles in Table 6.3 and the graphs in Figures 6.1- 6.4 suggested that $\bar{y}_{rsw}$ was conditionally biased. Empirical investigation suggested that $\bar{y}_{rsw}$ with sample variance adjusted by $(\bar{X}^2/\bar{x}^2)$ in (6.6) had better performance than that with the usual sample variance in (6.5) when the ratio estimate was employed with the agricultural survey cases. However, there was good evidence from Tables 6.4-6.5 suggesting that the post stratified mean $\bar{y}_{ps}$ was more efficient than $\bar{y}_{rsw}$ in these cases.

Chapters 7 and 8 introduced the distributions of the post stratified combined (separate) ratio. The percentiles in Tables 7.2-8.1 and the graphs in Figures 7.1-7.4 and 8.1- 8.4 indicate that $\bar{y}_{prc}$ and $\bar{y}_{prs}$ were conditionally unbiased. Empirical investigations suggested that $\bar{y}_{prc}$ and $\bar{y}_{prs}$ were more efficient than $\bar{y}_{rsw}$ in reducing the MSE as well as balancing out the conditional bias, whereas the self-weighting ratio estimate was very poor when the samples were unbalanced.

The conclusion to be drawn from the above results is that the confidence intervals using post stratified estimates $\bar{y}_{prc}$, $\bar{y}_{prs}$ and $\bar{y}_{ps}$ had approximately the correct coverage properties for each sample configuration obtained, and hence the correct coverage property over all possible samples, provided that the Central Limit Theorem was applied.

It also was concluded that the post stratified estimates $\bar{y}_{prc}$, $\bar{y}_{prs}$ and $\bar{y}_{ps}$ are (approximately) conditionally unbiased for the population mean, but the self-weighting estimates $\bar{y}_{rsw}$ and $\bar{y}_{rs}$ are conditionally biased.

In general, the post stratified estimates $\bar{y}_{prc}$, $\bar{y}_{prs}$ and $\bar{y}_{ps}$ are better than the self-weighting estimates $\bar{y}_{rsw}$ and $\bar{y}_{sw}$ when dealing with situations such as those illustrated by cases 3 and 4, and even when dealing with some artificial cases, such as in the cases 1 and 2 where stratification was efficient.

The sample variances $s^2_{\bar{y}_{sw}}$ and $s^2_{\bar{y}_r}$ represent the usual basis for estimating $V(\bar{y}_{sw})$ and $V(\bar{y}_{rsw})$ respectively, and our evidence shows that they may be misleading, and perhaps even inappropriate within a conditional framework. In particular, these sample variances overestimate the conditional MSE of $\bar{y}_{sw}$ and $\bar{y}_{rsw}$ for some configurations and underestimate them for others.

It was also concluded from these studies that employing conditional inferences with post stratification inflates the variances when the sample is badly balanced and reduces them when it is well balanced. Post stratification is a technique for protecting the statistician's inference against those occasions when his randomization gives an unbalanced or unrepresentative sample.

In this thesis we have used the case of an equal probability design where $\bar{y}_{sw}$ and $\bar{y}_{rsw}$ are the natural alternatives to $\bar{y}_{ps}$, $\bar{y}_{prc}$ and $\bar{y}_{prs}$. Our simulation results suggested that, if an auxiliary variable is available to construct strata, then $\bar{y}_{ps}$ is superior to $\bar{y}_{sw}$. Also in a situation where ratio or regression estimation was being considered but it was felt that the relationship might not be linear, post stratification could be viewed as a good alternative. From these simulations we conclude that, if the observed sample is disproportional and the variation from stratum to stratum is relatively constant, use $\bar{y}_{prc}$, but with disproportionate sample and unequal stratum variances, $\bar{y}_{prs}$ is recommended to use.

BIBLIOGRAPHY

Chester, Spencer, T. (1980). *The Use Of An Auxiliary Variable In Optimum Stratification.* Ph.D. thesis, Oregon State University.

Cochran. William G. (1977). *Sampling Techniques*, 3rd edition, Wiley, New York.

Dalenius, Tore (1950). The Problem of Optimum Stratification. *Skandinavisk*, 12, 351-356.

Dalenius, Tore and Gurney, Margaret (1951). *The Problem of Optimum Stratification. II. Skandinavisk Aktuarietidskrift*, 34, 133-148.

Dey Raj (1968). *Sampling Theory*, Mc Grow Hill, New York.

Fuller, W. A. (1966). Estimation Employing Post Strata. *Jornal of the American Statistical Association*, 61, 1172 - 1183.

Fuller, W. A. (1981). Comment on" An Empirical study of the Ratio Estimator and Estimators of its Variance," by R.M. Royall and W. G. Cumberland, *Jornal of the American Statistical Association*, 76, 78-80.

Holt, D. and Smith, T. M. F. (1979). Post Stratification. *Jornal of the Royal Statistical Society, A* 142, 33 - 46.

Kalton, G. (1984). Comment on " Present Position and Potential Developments : Some Personal Views sample surveys," by T. M. F. Smith, *Jornal of the Royal Statistical Society, A* 147, 220 - 221.

Kish, Leslie (1965). *Survey Sampling.* John Wiley & Sons, New York.

Robinson, J. (1987). Conditioning Ratio Estimates Under Simple Random Sampling. *American Statistical Association Jornal*, 82, 826 - 831.

Rao, J.N.K. (1985). Conditional Inference in Survey Sampling. *Survey methodolgy*,11,*No*.1,15 - 31.

Royall, R. M. and Cumberland. W. G. (1981). An Empirical Study Of the Ratio Estimator and Estimators of its variance. *Jornal of the American Statistical Association*, 76, 66 - 77.

Singh, Ravindra (1975). An Alternative Method of Stratification on the Auxiliary Variable. *Sankhya, Series C*, 37, 100-108.

Singh, Ravindra and Sukhatme, B. V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*, 21, 515-528.

Smith, T. M. F. (1984). Present Position and Potential Developments : Some Personal Views sample surveys. *Jornal of the Royal Statistical Society, A* 147, 220 - 221.

Thomsen, Ib (1976). A Comparison of Approximately Optimal Stratification given Proportional Allocation with other Methods of Stratification and allocation. *Metrika*, 23, 15-25.

Williams, W. H. (1962). The Variance Of An Estimator with Post Stratified Weighting. *American Statistical Association*, 57, 622 - 627.